Developing Transparency Requirements for the Operation of Criminal Justice Algorithms in New Zealand

Briony Blackmore

A thesis submitted for the degree of Master of Arts in Philosophy

The University of Otago
Dunedin
New Zealand
01/04/2019

Abstract

As predictive risk algorithms become more commonplace so have concerns about their use. The perceived bias, inaccuracy, and opacity of predictive risk algorithms has given rise to concerns about fairness when used in criminal justice contexts, especially in predicting an offender's risk of re-offense. Opacity in these algorithms has negative consequences for citizens' trust in their government, ability to give informed consent and ability to utilise certain rights. Edwards and Veale (2017) have argued that making predictive risk algorithms transparent would go some way toward mitigating these concerns. However, transparency is not a panacea: in fact, it introduces the possibility of offenders "gaming the system", thereby decreasing the potential accuracy and effectiveness—and indeed, the fairness—of the system. Complete transparency, then, does not guarantee fairness either. With this tension in mind, I explore the justifications for transparency and opacity in predictive risk algorithms. I argue that while neither complete transparency nor complete opacity is desirable, there are certain parts of predictive risk algorithms that are best kept opaque, while others should be transparent. I then propose a set of transparency conditions that should be met when operating algorithms in the criminal justice system.

Acknowledgments

Major thanks go to:

My supervisor Professor James Maclaurin at the University of Otago. His passion and excitement is contagious. Our meetings were always filled with laughter and great discussion. I would like to thank him for always motivating me to work harder (through excitement, not force), and constantly having thorough and clear advice.

My supervisor Professor Michael LeBuffe at the University of Otago. For stoically making his way through my final draft and for the thorough and careful comments. I would also like to thank him for always going above and beyond. It does not go unnoticed.

Joy Liddicoat, Dr. John Zerilli, Associate Professor Alistair Knott, and Associate Professor Colin Gavaghan at the University of Otago. For patiently, and clearly answering my many confused computer science, and law related questions (more than once). I have felt very lucky to have had a group of experts to talk to.

Dr. Peter Johnston, Director of Research and Analysis at the Department of Corrections. For being so open to answering my many questions in detail and for reading my final draft.

The Philosophy Department at the University of Otago, for creating an environment that was supportive and happy.

Joseph Burke, for the editing help, without which my thesis would be a far sight more tangled. Also, for the frequent coffee appreciation meetings. They were a highlight.

Daniel Jaffe, for always being up for an adventure in the mountains, out on bikes, or in the backcountry. There is no way my thesis would have been completed had I not been able to spend a lot of time outside not doing it.

My aunt, Penny Jones, for teaching me that I should love my work and for being a huge support in finding what it was that 'lit a fire' for me.

My parents, Susi and Tim Blackmore. For being my steadfast pillars of support, my biggest cheerleaders, and my greatest motivators. I also want to thank them for their immense help in proofreading my thesis, and for always encouraging me to dream big.

Table of Contents

Introduction

We see technology everywhere. You hardly ever come across someone who does not own a mobile phone, some cars are self-driving, and the public sector is using machine learning systems to predict the risk of certain events occurring. The uptake of machine learning, in particular, has increased substantially in a very short period and these systems are becoming extremely complex and sophisticated. In fact, many are far more capable than humans would be at specific tasks, such as making stock market predictions, and playing the ancient board game, Go (Gibney 2016, 445)[1]. From personally directed product advertising to university admissions, and from television show recommendations to decisions about insurance pay-outs, the application of this technology is both impressive and diverse. In various jurisdictions one use of machine learning tools is to calculate risk of recidivism scores. These scores mark how likely it is that an offender will reoffend within a certain time period. Originally an expert such as a psychologist completed this risk assessment but, as technology gets better, and more reliable, computer systems are also used to assess risk. There are multiple reasons that algorithms are useful in this context, including: increased accuracy, reduced bias, being able to "tweak" the algorithm and adding another source of information to risk assessments. One problem with these risk assessment algorithms is that they are often opaque. As a result, legal, technical, and ethical experts are calling for greater transparency as a result, but what exactly is required remains unclear (Kehl et al. 2017, 32, Rudin et al 2018, 1, Edwards and Veale 2017). This thesis is going to focus on algorithms that are used to predict the risk of a criminal offender re-offending (which from this point onwards will

[1] See McCaney (2018) for examples of other tasks where artificial intelligences are outperforming humans.

be referred to as criminal justice algorithms) and determine the degree and type of transparency required to build and operate such algorithms effectively and fairly.

Before going any further it is important to pause here and note that this thesis will not attempt to assess the use of the risk of recidivism as a factor in decision making. Whilst risk assessment looks as though it has a place in offering rehabilitation and parole, using it in sentencing decisions is contentious (Hope 2007, 1159). I will not be entering this debate in this thesis. This thesis is looking in particular at how best to operate criminal justice algorithms. It accepts that risk assessments are used and is looking at the most ethical way in which recidivism algorithms can be operated.

Operating an algorithm well through transparency

When operating an algorithm it is important to think about how to operate it well. That is, we need to be thinking about how to operate it in a manner that is ethical and fair. There is great discussion in the artificial intelligence literature about how best to do this (Select Committee On Artificial Intelligence 2018; Oswald et al. 2018; Tutt 2017). One way in which criminal justice algorithms might be operated well is to make how they work, how they are operated, and how they are used transparent to the public. However, to do this is not as simple as it sounds as there are several limits to transparency that must be considered. It is also important to make sure that transparency will not unduly compromise accuracy and fairness.

This thesis is going to argue that criminal justice algorithms should be transparent to such a degree that several things can be achieved. Firstly, so an offender give informed consent for their use. Secondly, so a data subject can appeal a decision that used a risk score produced by criminal justice algorithm. Thirdly, to keep developers and operators accountable. Finally, so the public can trust the use of an algorithmic system. It will also

argue that depending on the algorithm used there are limits to transparency that must be recognized and observed in order to avoid compromises to accuracy and fairness.

Structure of thesis

In order to argue that there needs to be transparency in order to operate criminal justice algorithms well a lot of groundwork must be done. The first chapter will introduce criminal justice algorithms and how they are used in criminal justice systems around the world. It will then justify the use of criminal justice algorithms alongside the use of human risk assessment by arguing that algorithms are more accurate, have the potential to be less biased, are "tweakable", and finally can be a helpful additional information source. Three criminal justice algorithms will be introduced and they will be used as case studies throughout the remainder of the thesis. The chapter will then turn to the issue of transparency and suggest that in order to operate criminal justice algorithms well they should be made transparent. It will also discuss briefly some reasons for why transparency might help algorithms operate well.

To further set the scene, chapter two will explain what criminal justice algorithms are. Not all criminal justice algorithms use the same model to make predictions and each of these models poses different practical problems when it comes to being able to provide explanations for how algorithms work. It is important, then, to understand how algorithms can differ and how this can impose limits on transparency. This chapter will begin by discussing the history of risk assessment and will explain how it has changed over time. It will give an in-depth explanation of what an algorithm is and how it is built. From here, using the case studies introduced in the first chapter, I discuss different models of algorithms used to make predictions and the benefits and weaknesses of using certain models.

In chapter three the issues of accountability, informed consent, rights of appeal, and public trust (which are briefly cited in chapter one as reasons for transparency) will be discussed in full. Arguments will be given for why accountability, informed consent, rights to appeal and public trust are important elements of a well-operated algorithm. Having established this I will argue that transparency is way of achieving these elements. I will explain what degree and type of transparency is necessary.

In chapter four I will introduce four justifications for criminal justice algorithms remaining opaque in some way. These justifications are intellectual property, gaming the system, reducing public trust, and the complexity of algorithms. I will explain all of these justifications in full. I will then assess whether any of these justifications need to be taken into consideration when it comes to regulating the level of transparency needed to operate criminal justice algorithms well. In cases where there look to be tensions with reasons for transparency I will make suggestions about how they might be mitigated

Chapter five puts the arguments in chapter three and four into practice. In this chapter I will argue that the best way to regulate is to have an independent specialist set regulations for the use of criminal justice algorithms. I will also argue that attempts at forming regulations thus far have not been appropriate and have not managed to ensure the transparency needed to operate criminal justice algorithms well. Then, taking the outcomes of the arguments presented in chapters three and four, I will make suggestions for how best to regulate transparency in criminal justice algorithms within the context of New Zealand's criminal justice system.

1.0. Chapter One: What criminal justice algorithms are, why we should use them, and justifications for transparency

1.1. Introduction

In multiple jurisdictions around the world predictions are made about how likely it is that a criminal offender will reoffend. One way of assessing this is to use a predictive risk algorithm, and many jurisdictions use this approach. This first chapter will serve as an introduction to where and how criminal justice algorithms are used to predict risk of recidivism. In section 1.2. I will explain what a criminal justice algorithm is. In section 1.3. I will explain why criminal justice algorithms are used in the criminal justice system and discuss what they add to a human's judgment of risk. In section 1.4. I will introduce three specific criminal justice algorithms used around the world and these will then be used as examples throughout the remainder of this thesis. In section 1.5. I will discuss motivations for focusing on transparency and algorithms in this thesis. Furthermore, I will discuss what transparency and opacity mean in the algorithmic context. In section 1.6. I will introduce common justifications for keeping how an algorithm works and how it is used transparent. Finally, in section 1.7 I will discuss the requirements for transparency between an operator and a developer of an algorithm.

1.2. How criminal justice algorithms are being used in criminal justice systems

Criminal justice algorithms are being used in various jurisdictions to predict the risk of an offender re-offending, whether a reoffence will be violent, and in some cases, whether a re-offence will lead to imprisonment (Nadesu 2007, 15). It is well known that someone who has committed a crime is more likely than a member of the general population to enter the judicial system in the future (National Institute of Justice 2014). In fact, 49% of all released

prisoners in New Zealand are re-imprisoned within just five years of their release (Stats NZ 2018, 21). It is therefore useful to assess how likely it is that each individual offender will re-offend. Such assessments provide corrections staff with useful information that can help them make decisions about future prison costs, which rehabilitative measures and techniques an offender should receive, whether an offender should receive bail or parole, and in some jurisdictions, about an offender's sentence (Eaglin 2017, 61). Assessing the risk of an offender re-offending is not new. This has been done for many years. However, it used to be done solely by means of human judgment (Eaglin 2017, 67). Psychologists, case workers, and prison workers used to offer their educated opinions, which would then be used to help make decisions that would affect an offender. Now, the opinions of these experts are still used, but there is an additional information source: a risk score that is developed by a criminal justice algorithm.

Criminal justice algorithms use several facts about an offender to calculate a risk score. Different systems use different facts, but they all include variables that have been found to be closely correlated to those who reoffend, such as an offender's criminal history and the seriousness of their previous offence(s) (Eaglin 2017, 69, 84). Some criminal justice algorithms also use facts that come from asking an offender their thoughts on certain aspects of their life, such as their friendships, or how often they feel bored (Northpointe 2016). The variables are used as inputs to the criminal justice algorithm. The criminal justice algorithm then completes a process (which differs depending on the predictive model used[2]) and a numerical risk score is produced as an output. This numerical risk score indicates the probability of an offender re-offending within a certain period following their offence and is based 'on the behavior of the individuals in the underlying data set' (Eaglin 2017, 85). Public

[2] The primary predictive methods used in criminal re-offense prediction will be discussed in §2.5.

and private entities work to assign different levels of risk to the scores that a criminal justice algorithm produces, which is a highly subjective, policy related process (Eaglin 2017, 85, 87) (See Appendix A for the RoC*RoI risk level table). That is to say policy makers decide what range of probabilities should be considered low risk, moderate risk, and high risk. For example, the criminal justice algorithm might determine that an offender has a 17% chance of re-offending. If the government agency has decided that anything under 25% is categorized as low risk, then the offender is labelled as being at low risk of re-offence and is treated accordingly. This risk score does not decide the fate of an offender, rather the risk score is used, alongside other information to make decisions about that offender (Stats NZ 2018, 9). In the *Algorithm Assessment Report* (2018), which provides a stock take of government operated algorithms, Stats NZ have suggested that it stay this way, and that we should not allow an algorithm to make the final decision about a data subject (a data subject is the actor who is impacted by the output of an algorithm) (2018, 9). Furthermore, Oswald et al. suggest that in cases where difficult risk assessments occur, 'a fair and trustworthy algorithmic decision-making tool may potentially be helpful, provided not used in a determinative way'(2018, 225).

1.3. What do algorithms add to risk assessment?

We might ask what a criminal justice algorithm offers that a human who assesses risk does not. To justify the use of a criminal justice algorithm in the criminal justice system it is important to explain what the use of an algorithm adds to risk prediction (Kehl et al. 2017, 34).

1.3.1 Accuracy

When determining the risk of an offender committing another crime we want the assessment to put the offender in the risk category that best reflects the actual risk that they are at of re-

offending. The reason this is important is that it would be bad to treat someone who is low risk as high risk, and vice versa. It is unlikely that we will ever get to a point where we accurately capture every single person's level of risk. However, it is important to make sure that offenders' levels of recidivism risk are captured correctly as often as possible as otherwise we might send many people to prison who should not be there, or let people out on parole who then go on commit a serious crime. It is good to decrease the number of times that an offender is predicted to be high risk of reoffence and they never offend again (false positive), and the number of times it is predicted that an offender is low risk and they offend again, perhaps violently (false negative) (Dressel and Farid 2018, 1).

Criminal justice algorithms are thought to be better at predicting whether an offender will offend than even a well-prepared human-being is at making the same prediction. It is with this thought in mind that police and law enforcement agencies are beginning to use this technology (Zerilli et al. 2018, 2). Whether criminal justice algorithms are in fact more accurate at making such predictions as humans are is largely is an empirical question. Data on this is hard to come by, but it is accepted that actuarial assessments largely have superior accuracy compared to assessments made by human-beings (Craig and Beech 2009, 206; Fry 2018, 55)[3]. New Zealand's Algorithm Assessment Report also reports, that 'even simple risk scales (ie [sic] a checklist of risk factors) invariably outperform the clinical or professional judgments of trained experts and experienced correctional staff when making predictions about future offending' (Stats NZ 2018, 21). Advocates of algorithm use in criminal justice systems also argue that criminal justice algorithms are more objective and consistent in their

[3] Actuarial Risk Assessments are risk assessments that are 'largely focused on historical or unchanging risk factors. When an actuarial instrument is used to assess risk, an offender is scored on a series of items that were most strongly associated with recidivism in the development sample. The offender's total score is cross-referenced with an actuarial table that translates the score into an estimate of risk over a specified time frame.' (Desmarais and Singh 2013, 5)

decision making compared to a human decision maker and as a result using the tools will lead to more accuracy (Eaglin 2017, 62; Fry 2018, 54). If a system is more accurate than a human being at predicting reoffence and it does not pose big ethical problems, then it should be used.

1.3.2 Bias

People are innately biased. They want to help those that they are close to or find similarities with and sometimes disfavour those that they are dissimilar to or do not like (Bloom 2016, 68; Hoffman 2000, 13–14). This is problematic if it motivates action without good reason in the justice system. Decisions about an offender's journey through the judicial system should not be based on personal, or group biases. The hope is that by using a mechanical system, such as an algorithm, to predict risk it can remove the biases that inevitably make their way into assessments completed by humans. However, algorithms currently in use are not free of biases. There are several reasons why algorithms will produce biased results. Firstly, human beings construct algorithms. This means that the developer of an algorithm can weight contributing factors in a manner that reflects a bias towards particular groups of people. COMPAS, a criminal justice algorithm used in various part of the United States, has come under fire for exactly this, with Larson et al., from ProPublica, claiming that COMPAS is biased against black people (Larson et al. 2016)[4]. Bias could be completely unintentional, or intentional, on the developer's part (Eaglin 2017, 97). Policy makers that work alongside developers could spread their biases by pushing for certain factors to be used as variables in an algorithm. They could also contribute their own biases when determining the parameters of the low, moderate, and high risk bands (See appendix A for the RoC*RoI risk bands).

[4] Whether this bias actually exists, it is difficult to tell. Larson, et al. (2016) of ProPublica claim that it does, whilst Dieterich, Mendoza, and Brennan (2016) claim that it does not. For more; read (Larson et al. 2016) and (Dieterich, Mendoza, and Brennan 2016)

Secondly, biases can stem from the historical data that is used to construct an algorithm and no predictive tool can be better than the data from which it is developed (Eaglin 2017, 72). As will be explained in §2.5, algorithms that calculate risk of recidivism are built using the data of previous offenders in that jurisdiction. If that historical data includes biases towards groups, so will the criminal justice algorithm that is built. Developers might tweak the algorithm in order to avoid such bias of course, but in doing so they need to be careful not to introduce their own biases (Fry 2018, 69). Hopefully over time, biases can be removed from these systems and if this happens then the algorithms will act in an unbiased manner with greater consistency than a human is able to.

1.3.3 Algorithms are more 'tweakable' than humans

Hannah Fry, a mathematician at University College London, makes a strong case for algorithm use by explaining that an algorithm can be tweaked in in a way that humans cannot (Segal 2018). If a human risk assessor, such as a psychologist or caseworker, is making inaccurate, biased, or unfair decisions about an offender, they can be told to change their behaviour, which may or may not lead to a change. Even when a person is aware that they are being biased in their decision-making or doing an inaccurate job in their assessments, they may not be able or may not want to change their decision-making process to correct their fault. On the other hand, as Fry points out, the strength of a risk prediction algorithm is that, if it is noticed that an algorithm is making inaccurate or biased predictions there is a way to change this by 'tweaking numbers accordingly', that is to say, by re-weighting inputs (Fry 2018, 69). Whilst it might be extremely difficult to determine what needs altering, once done, it will change the behaviour of the algorithm and this can be done instantly, or over time (Fry 2018, 70).

1.3.4 Criminal justice algorithms bring a new source of information to decision making

When making decisions that have an immense impact on people, such as whether an offender receives parole, decisions should be well-informed ones. A criminal justice algorithm is an additional and different type of risk assessment source to psychologists and caseworkers. To help judges, parole board members, and corrections staff make better informed decisions about offenders adding an additional source of information is beneficial if it is as good or better than human judgment. One might object and say that a person is able to make the calculations that a criminal justice algorithm can. However, I would claim here that there is a difference in kind, in terms of the abilities of the human and the criminal justice algorithm. A criminal justice algorithm can deal with a lot more data, and more efficiently. In addition, a criminal justice algorithm, can be altered to be less biased or inaccurate, in a way that a human cannot (Fry 2018, 65; Segal 2018).

1.4 New Zealand, United States of America, and United Kingdom case studies

There are many criminal justice algorithms used around the world to assess the risk of an offender re-offending. There are too many to focus on, so instead I will discuss three in depth, and then use these as examples throughout this thesis to help develop my recommendations for transparency. I have chosen the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm developed in the United States, the RoC*RoI (Risk of Recidivism x Risk of Imprisonment) algorithm developed and used by New Zealand's Department of Corrections, and the HART (Harm Assessment Risk Tool) algorithm, used by County Durham's Constabulary in the United Kingdom. These criminal justice algorithms differ in several ways; the ways in which they are used in their corresponding jurisdictions, the predictive models they use, the variables they use, as well as how transparent their

developers and operators are with data subjects, and the public. This makes them useful for comparative purposes.

*1.4.1 RoC*RoI Algorithm*

RoC*RoI is a sophisticated computerised risk prediction algorithm developed by the New Zealand Department of Corrections, and has been used to assess risk of re-offense and imprisonment since 2001 (Nadesu 2007, 15). The algorithm produces a risk score for each offender. Corrections staff and the parole board then use this risk score, in addition to a psychologist's report and other expert opinions, to make decisions about an offender (Stats NZ, 21). These decisions relate to: the level of management required on community based sentences, offender eligibility for rehabilitation programmes, parole release, as well as a prisoner's security classification (Stats NZ 2018, 21). The RoC*RoI algorithm calculates both the risk of re-offence and the risk of imprisonment, and these scores are multiplied together to produce the RoC*RoI score (Department of Corrections New Zealand 1999). This score sits within a certain 'band' of risk, within which 'a certain proportion are expected to reoffend' (Johnston 2019) (Refer to Appendix A to see the risk band classification used by the New Zealand Department of Corrections). There are a number of variables that are used for each of these assessments (a full list can be found in Appendix B (Johnston 2018). Some of these variables include; age, sex, frequency of offence, severity of crime, and time spent in prison (Stats NZ 2018, 21). They are static factors (explained in §2.3) and are facts an offender cannot change about themselves. These RoC*RoI scores are developed routinely and the offender does not need to give consent for the Department of Corrections to develop the score (Johnston 2018).

1.4.2 COMPAS algorithm

COMPAS is a criminal justice algorithm used in multiple jurisdictions in the United States. COMPAS was developed by a company called Northpointe Institute for Public Management Inc. (now called Equivant) and is sold to agencies throughout the United States that add their own data to the algorithm (Eaglin 2017, 69)[5]. Each jurisdiction can choose how they use the risk scores that are developed by the COMPAS algorithm. Some states, such as Michigan, use it to help guide decisions about programmes a prisoner might be included in, rehabilitation, and parole (Michigan Department of Corrections 2017, 4). Some states, such as Wisconsin, also use the score produced by the algorithm to make decisions about an offender's sentence (Freeman 2016, 75). The COMPAS algorithm uses 137 variables to produce a risk score (Dressel and Farid 2018, 1). The algorithm uses both static and dynamic variables to develop a risk score (Northpointe 2015, 1). This means, they use both unchangeable facts about the offender (static factors), such as their criminal history, and changeable details (dynamic factors), such as how often an offender feels bored, or whether they consider themselves to have close friends (Northpointe 2016) (A copy of the 137 questions used by the COMPAS algorithm is available in appendix C).

1.4.3 HART algorithm

The HART algorithm was developed by experts in statistics at the University of Cambridge, in collaboration with Durham Constabulary (Oswald et al. 2018, 225). It is operated by the County Durham Constabulary in the United Kingdom. The Durham Constabulary use this algorithm to determine the risk of an offender re-offending. Offenders are grouped into high, moderate, and low risk of re-offence. High risk offenders are those classed as likely to

[5] Northpointe Inc. has changed their name to Equivant (Bennett Moses and Chan 2018, 261). Throughout this thesis I will use Northpointe as the majority of literature is written in reference to Northpointe.

violently reoffend in the following two years, those at moderate risk are classed as likely to reoffend, but not violently, and low risk are predicted not to reoffend in the following two years (Oswald et al. 2018, 227). Those who are classed as at moderate risk of re-offence are permitted to be part of the constabulary's Checkpoint Programme which has been developed to try and tackle root causes of offending (Oswald et al. 2018, 227). Instead of sending all offenders through the prosecution process, those thought to be at moderate risk of re-offence, and thought able to be rehabilitated, are admitted to the Checkpoint programme which provides them with an alternative to court and prosecution (Oswald et al. 2018, 225). Whilst HART operators and developers have not made a list of variables available, they state that they use 34 variables as inputs (Oswald et al. 2018, 228). These variables, are largely based on the criminal offence history of the offender (Oswald et al. 2018, 228). Additional variables include; age, gender, postcode, and the number of already existing police intelligence reports on the offender (Oswald et al. 2018, 228).

1.5 Transparency and criminal justice algorithms

1.5.1 Defining Transparency

The language surrounding transparency and algorithms is difficult to navigate and is ambiguous at best. We use the terms "transparency" and "opacity" in relation to algorithms often, but, for the most part, these terms are not properly defined. We can understand complete transparency as being able to access the source code for an algorithm, having access to information about how an algorithm works, knowing what factors are being used, and knowing what an algorithm is being used for. On the other hand, we can understand complete opacity as the antithesis. There would be no access to source code, no information about how an algorithm works, or of the factors being used, and no knowledge that an algorithm is even being used. We are unlikely to ever ethically require complete opacity or complete

transparency for these systems, but it is helpful to see the bounds we are working within. The recommendations that I will make throughout this thesis will be explicit about the degree of transparency required, as well as when throughout an algorithm's use it is needed. The transparency requirements I suggest regarding how and algorithm works and how it is used will also come in two different forms; *ex ante*, and *ex post* transparency. I count *ex ante* transparency as details given prior to an event, and *ex post* transparency as details given during or after an event. *Ex ante* transparency in relation to an algorithm will be understood to be transparency about an algorithm prior to its use, and *ex post* transparency, to be transparency after an algorithm is used.

1.5.2 Motivation for focusing on transparency

As discussed in the extended introduction, this thesis will focus on transparency. Criminal justice algorithms need to be operated well. One way in which a criminal justice algorithm can be operated well is by making elements of how it works and how it is operated available to those affected by it. However, the solution is not to make the entire system transparent. This thesis will identify the ways in which the algorithm and its use must be made transparent as well as when it is not possible for it to be transparent. In doing so, recommendations for regulating criminal justice algorithms can be developed.

1.6 Reasons for making algorithms transparent

In this section I will discuss several justifications for disclosing as many details about the algorithmic system as possible. It is important to note here that I am not engaging in arguments about whether these justifications are good and whether they require transparency. That will be the focus of chapter three.

1.6.1 Trusting criminal justice algorithms

The public is worried about bias and accuracy in criminal justice algorithms, which is somewhat ironic given that the aim of using an algorithm is to help us be less biased and more accurate in our decision-making. The public wants to know, as with human made decisions, that the decisions are both fair, and reasonable (Zerilli et al. 2018, 5). They also want to know that the tools are really doing what the operators and developers say they are doing (Eaglin 2017, 106). They want more than just reassurance, they want to see how the system works; what factors are used, and how the inputs produce the output. On face value, it is not entirely unreasonable that the public might desire this, particularly as studies such as the ProPublica analysis of COMPAS have reported that the COMPAS algorithm is biased towards white people and against black people (Larson et al. 2016).

If the public is unable to access information that they desire about how a criminal justice algorithm works, public distrust in the criminal justice system may develop. Democratic trust is important, a democracy thrives on some level of it and, in fact, it is often considered a 'necessary precondition for democratic rule' (van der Meer 2018, 1). Keeping algorithms opaque and not giving citizens information about their inner workings may well lead to decreased trust, and in this sense it may be in the best interest of government departments to be more transparent with citizens.

Not only does a democratic government thrive on trust but as stated in New Zealand Statistics' *Algorithm Assessment Report* '[i]f the process is not transparent, people are less likely to trust the decisions that are made as a result, and may not wish to continue sharing their personal information' (Stats NZ 2018, 7). Not sharing information then has the knock-on effect of the criminal justice algorithm having less data to develop risk scores from, which in turn likely decreases accuracy.

1.6.2 Accountability

Relevant to the justification of trust is the justification of accountability. As van der Meer states: a person is likely to lose trust when they are unaware of who is accountable. Assigning accountability to various actors in the development and operation of a criminal justice algorithm might help to ensure that it is as unbiased and accurate as possible and motivate operators to use it in the most ethical way possible (Mulgan 2000, 556).

The question becomes: what relevance does the discussion of accountability have to transparency? Statistics New Zealand's Algorithm Assessment Report states, that maintaining transparency is in fact 'essential for accountability' (Stats NZ 2018, 9). If this really is the case, and we do require accountability, it looks as though transparency must also be required to some degree.

1.6.3 Appealing a decision, judicial review, and statutory rights

As Zerilli et al. point out, for an offender to appeal decisions that are made about them in the justice system, they need to know the basis on which a decision has been made (2018, 4). If an algorithmic system hides information that demonstrates the basis on which a decision is made then an offender does not have appropriate information available for them to appeal a decision, ask for a judicial review, or exercise a statutory right to review. If a jurisdiction gives an offender these rights, then a justification for transparency might be that there needs to be an opportunity for the offender to have access to information that will be grounds for an appeal. Furthermore, an offender should receive an explanation for why an appeal was rejected or accepted, and this might include transparency about the algorithm itself, or how the algorithm was used.

1.6.4 Informed consent

The concept of informed consent is most commonly found in medicine. A patient must give a medical professional consent before a procedure is carried out in most cases (Eyal 2019). To give consent, the patient must fully understand what they are consenting to, though it is difficult to articulate what 'fully understanding' amounts to.

A justification for making the inner workings of a criminal justice algorithm available to an offender is that to develop a risk score, an operator might require informed consent from the offender; much like consent is required from a patient about to undergo a medical procedure. The New Zealand Department of Corrections does not require consent from a prisoner for the RoC*RoI algorithm to produce a risk score (Johnston 2018). On the other hand, criminal justice algorithms such as COMPAS require that the defendant provide the information voluntarily, which is to say that they can choose not to provide information for a risk score to be produced (Eaglin 2017, 85). This raises questions about both whether an offender ought to give consent, and whether an opaque algorithm is a roadblock in terms of giving consent if it is required.

If it is found that an offender must consent to the use of a criminal justice algorithm then it looks like transparency to some extent may be essential. In order for the consent to be "informed" the operator of the algorithm would need to disclose some details to an offender prior to the risk score being developed.

1.7 Transparency requirements between operators and developers

Most of the discussion about transparency and criminal justice algorithms is centred on the information that data subjects and the public should receive. Discussion about the transparency required between developers and operators is also important, but often falls by

the wayside. In this section I will discuss the transparency required between operators and developers and the benefits that transparency between these two actors can provide.

There should be no restrictions on transparency between a developer and an operator of a criminal justice algorithm. This is not to say that there is a requirement for "complete transparency" but rather that requests for information from either party should not be refused. The benefits of doing this are two-fold. Firstly, transparency can act as an enabler in this context (Wanna 2018, 12). What I mean here is that transparency between an operator and a developer can lead to algorithms that are better constructed for their purpose, as well as to better developed policy and practice. Transparency would help achieve this as 'everyone has access to the information on which decisions are based and the assumptions informing decisions' (Wanna 2018, 12). Secondly, transparency combined with education can lead to better explanations for data subjects and the public. Developers and operators of algorithms have largely different skillsets. Developers are skilled in statistics and computer science, whereas operators are skilled in public policy and government operations. It is hardly surprising then, that developers often do not fully understand the operation of the algorithm, nor do the operators understand how the algorithm works. It would be pointless to require that each completely understood how the others realm worked, as it is unlikely that either party could achieve this. However, by combining unrestricted sharing of information between developer and operator and requiring that these actors educate one another, better understanding can be achieved. This may have the outcome of being able to provide better explanations for data subjects and citizens.

This section has stated that un-restricted transparency between operators and developers is important for two reasons a) to develop better algorithms, better practice, and better policy and; b) to provide better understanding and therefore better explanations for data subjects and citizens. Having this transparency requirement between developers and

operators stands the operators and developers in good stead for being able to provide the transparency I determine is required in chapters three and four.

1.8 Conclusion

In this chapter I have explained that algorithms are used in many jurisdictions to produce a risk score that indicates the level of risk that an offender is at of re-offending. This risk score is then used to make decisions about the offender in relation to rehabilitation, parole, and in some cases sentencing. I have then given four reasons for which a criminal justice algorithm might be a welcome addition to risk prediction in the criminal justice system: increased accuracy, less bias, a tweakable form of risk assessment, and an additional source of information. I have also introduced three algorithms that will serve as case studies throughout this thesis. These algorithms are New Zealand's RoC*ROI algorithm, Northpointe's COMPAS algorithm, and County Durham's HART algorithm. Having explained these, I introduced the concept of transparency which is the main focus of this thesis. I defined transparency, and explained that this thesis will focus on transparency as it appears to be important when it comes to operating a criminal justice algorithm well. I then briefly discussed four reasons that transparency is important for operating an algorithm well. These reasons included trusting criminal justice algorithms, accountability, rights to appeal criminal justice algorithms, and informed consent. I finish this chapter by outlining the level of transparency required between developers and operators to ensure that important information for data subjects and citizens can be conveyed to them.

In the chapters that follow I will explain how algorithms used in the criminal justice system work, whether the justification in sections 1.6.1. – 1.6.4. are reasons to require transparency, and whether there are restrictions to transparency. Having done this, in my

final chapter I will construct a framework for regulating the transparency of criminal justice algorithms in the New Zealand criminal justice system.

2.0. Chapter Two: Decoding Predictive Algorithms

2.1. Introduction

A variety of risk assessment tools are used for different purposes in criminal justice systems around the world. Each of these tools present a challenge when it comes to explaining how they develop a risk score. This can differ depending on the type of model used. In this chapter I shall explain how risk assessment algorithms work. Firstly, I will give an overview of the generations of risk prediction tools, and then I will look at a series of important concepts relevant to criminal justice risk prediction such as dynamic and static factors, and supervised and unsupervised learning. I will then explain how random forest, regression analysis, and artificial neural networks work as these are a representative set of models used in criminal justice risk prediction.

2.2. Generations of risk assessment tools

There are multiple generations of risk prediction tools. Originally risk prediction was carried out solely by human-beings making judgment (Desmarais and Singh 2013, 4). However, developments in technology have added another sort of risk assessment to many jurisdictions. These are risk assessments done by algorithms

First generation recidivism prediction is essentially 'unstructured professional judgment' (Desmarais and Singh 2013, 4). There is no set checklist or protocol for risk assessment and essentially the assessor gathers information from the offender as well as from their official records, in order to make their judgment (Gottfredson and Moriaty 2006, 180). This was a very common form of risk assessment in the 1970's (Desmarais and Singh 2013, 4). An explanation of how the decision was made would come directly from the person who made the decision. However, with technology taking over many tasks previously entirely

completed by humans, this is a less common form of risk assessment now (Desmarais and Singh 2013, 4).

Second generation tools are comprised of historical and static factors such as criminal offence history and are actuarial in nature (Desmarais and Singh 2013, 4)[6]. These tools guide assessors to 'consider a set list of risk factors to arrive at a numerical risk of recidivism' (Desmarais and Singh 2013, 4). However, with the discovery that dynamic factors such as substance abuse, and mood help to increase the accuracy of risk prediction, second generation tools have become increasingly less popular (Desmarais and Singh 2013, 4)[7].

Third generation risk assessments are characterised by tools that incorporate dynamic factors and criminogenic needs, and 'may use an actuarial or structured professional judgment approach' (Desmarais and Singh 2013, 5). Static factors are still often used in these models, but the addition of dynamic factors, such as attitudes and substance abuse, may lead to changes in 'risk levels over time and can assist in identification of treatment targets' (Desmarais and Singh 2013, 5).

Fourth generation models also incorporate both static and dynamic factors. A difference between third and fourth generation tools is that fourth generation tools 'integrate case planning and risk management into the assessment process' (Desmarais and Singh 2013, 5). These models, are not simply looking at assessing risk, but also focus on 'enhancing treatment and supervision' (Desmarais and Singh 2013, 5). This method also allows for 'the role of professional judgment while remaining grounded in research and theory' (Desmarais and Singh 2013, 5). In this thesis, I shall be focusing on the computer driven algorithms that fall between the second and fourth generation of risk tool.

[6] For more information on actuarial risk assessments refer to footnote 3.
[7] I explain and assess dynamic factors in section 3.3.

2.3. Static and dynamic factors

As I stated in the previous section, early iterations of risk prediction tool included only static factors for recidivism. However, later generations have started to use dynamic factors. There are notable strengths and weaknesses for both methods. Static factors are factors that are 'historical or otherwise unchangeable' by an offender, such as criminal history and age which help to establish an 'absolute level of risk' (Desmarais and Singh 2013; Ward and Fortune 2016, 80). New Zealand's RoC*RoI algorithm is an example of a criminal justice algorithm that uses static factors to generate a risk score (Stats NZ 2018, 21). Dynamic factors, on the other hand, are factors that can change and include things such as substance abuse and antisocial behavior and thoughts (Desmarais and Singh 2013). These dynamic factors help to determine a relative level of risk and can either be relatively stable and slow changing, or acute (Desmarais and Singh 2013). Acute dynamic risk factors are 'highly transient conditions that would only last hours or days' (Hanson et al. 2007, i). Stable dynamic risk factors are those that are 'personal skill deficits and predilections and learned behaviours' that are correlated with reoffence (Hanson et al. 2007, i). The COMPAS criminal justice algorithm developed by Northpointe, and used in many United States jurisdictions, is an example of an algorithm that uses both static and dynamic factors to generate a risk score (Northpointe 2012, 1). Third and fourth generation risk prediction methods incorporate both static and dynamic factors, and some developers claim that this approach leads to more accurate prediction (Desmarais and Singh 2013, 8). This approach however, is not without problems.

There are two main weaknesses when it comes to using solely static factors. One problem is that they identify unchangeable features which means that they 'do not offer opportunities for intervention' (California Coalition on Sexual Offending 2015, 3). That is to say, they do not identify areas in which rehabilitation efforts could be focused, such as

drug dependence, or violent behaviour (Gendreau et al. 2012). Secondly, whilst using static factors to predict recidivism gives a reasonably accurate estimate of long term risk, they give no hint about particulars, that is, which of these people will be reoffenders or, when re-offence may occur (California Coalition on Sexual Offending 2015, 3).

One reason for including dynamic factors is that predictions about an offender's risk of reoffending tend to be more accurate than simply using static factor assessments (Beech et al. 2002, 155). Furthermore, whilst some are wary about whether (and how) dynamic factors should be used in risk assessment, a distinct advantage is that they identify areas in which rehabilitation efforts should occur (Ward and Fortune 2016, 80). According to Northpointe using dynamic factors also allows 'for measures to change over time as behavior changes' as well as allowing for an '"overlay" of previous assessments on the latest assessment' (Northpointe 2012, 1). As Ward and Fortune point out, their use has also resulted in the development 'of intervention programs design to modify the characteristics of individuals and their environments associated with crime' (Ward and Fortune 2016, 80). COMPAS, used in several states in the United States of America, is a tool that uses dynamic factors as inputs in their algorithms[8]. It is useful to approach the use of dynamic factors with trepidation though, as it could be morally problematic to make predictions based on an offender's feelings, or attitudes about something, rather than what they have done[9].

[8] Dynamic factors are often viewed as both useful for predicting re-offence, and also potential causes of reoffending. However, Ward and Fortune disagree. For more a more in depth description, read Ward and Fortune 2016.

[9] This is not intended to turn into a discussion of whether it is morally appropriate to use dynamic factors. Rather it is to explain that they are controversial but there are strengths to using them.

2.4. What is an algorithm?

In this thesis, I am focusing on the second to fourth generation of risk prediction tool, specifically those that produce a prediction about the risk of an offender reoffending via a computer based algorithm. I am interested in algorithmic systems in particular because they face objections about their 'opacity' (Kehl et al. 2017, 28). What I mean by this is that people object to their use, as they cannot infer information about how the algorithmic system works and how it determines a risk score from the model. Before I explain the various types of models that may be used for criminal justice algorithms I need to explain what an algorithm is. According to Coffin and Saltzman an algorithm is 'a description of a mechanical procedure for carrying out a computational task' (Coffin and Saltzman 2000, 24). Once the model has received the inputs, essentially, it is, 'a sequence of instructions that are carried out to transform the input to the output' (Alpaydin 2016, 16). So, in the case of criminal justice algorithms, a prisoner's answers to a questionnaire are fed into the model as inputs and the algorithm produces (as an output) a risk of recidivism score through following a sequence of instructions.

2.5. Algorithms and how they "learn"

Algorithms used to determine an offender's risk of recidivism can learn (Curtis 2018, 2). They do this by detecting patterns in data (Kim 2017, 2). We can liken this to our own learning; after many attempts at catching a ball, we would be able to solve the 'catching problem', that is, we would be able to catch a ball. We would do this by developing an internal model of the parabolic flight of a ball. We would call this learning the skill of catching a ball. Algorithms do similarly, after looking at enough data they can solve a 'prediction problem' by developing a pattern from the data. In time a criminal justice algorithm would acquire the skill of predicting the risk of an offender reoffending.

Learning algorithms can be developed either through supervised or unsupervised learning. Criminal justice algorithms are supervised learning models (Curtis 2018, 2). Predictive risk algorithms that are used in criminal justice systems learn in a supervised environment. Supervised learning occurs when an algorithm has been trained in a domain where the output is known in advance, so it is known when the algorithm gives what counts as the right answer. An example of this type of learning is training a facial recognition algorithm where we know who a face belongs to. When the model claims that a photo of Angus is Angus we can know it is working because we know what Angus' face looks likes.

It is important that supervised learning occurs in the criminal justice case because it is necessary to check that a criminal justice algorithm is making predictions that accurately reflect the outcomes. By this, I mean it is important that the developer of an algorithm knows whether an algorithm is producing reasonably accurate risk scores from the answers that the offender gave to the questionnaire they completed, or the facts used about them. The reason that this is important is that the outputs are used in high stakes situations such as parole and sentencing decisions.

When an algorithm is being trained, the model is provided with both input and output data. Testing is then completed to check the system by using some withheld training data (LeCun, Bengio, and Hinton 2015, 437). The input training data is fed into the system and then the withheld output training data is compared to the output that the model produced from the input data (Kim 2017, 13). This process tests how well an algorithm predicts. In the criminal justice algorithm case, when training the model, the developer knows the questions that the offender was asked and also knows whether that offender reoffended in a certain time period.

Unsupervised learning, on the other hand, is best used when you are trying to work out a pattern in some data. With unsupervised learning you do not know the answers in

advance, but are rather using an algorithm to explore a data set in order to find useful patterns such as clusters of cases that are similar. Here the algorithm takes all the input data and develops a pattern from this (Kim 2017, 13). Algorithms of this sort are not (yet) used in the criminal justice system. An example of a model that uses unsupervised learning is a case where an artificial neural network identified dementia by analyzing the patterns in many EEGs (Electro Electroencephalogram) (Baxt 1995, 1137)[10].

2.6. Different models used to develop algorithms used for risk prediction

There are several different ways in which machine learning can help to forecast the risk of recidivism and different criminal justice algorithms used around the world utilise different methods. There are two main types of machine learning known as statistical methods and deep learning methods respectively. Statistical modelling methods are a 'formalization of relationships between variables in the form of mathematical equations' (Srivastava 2015). For example, a statistical model could formalise the growth of a butterfly population. It is important to note here that there are statistical methods that do not learn; however, in the case of statistical predictive risk models used in the criminal justice system the statistical models used are, in fact, supervised learning models (Curtis 2018, 2). Deep learning algorithms on the other hand are models that lack rules-based programming, in the same way that human brains do not have rules hard coded into them (Srivastava 2015). Currently, criminal justice algorithms such as COMPAS, RoC*RoI and HART use statistical methods of prediction, but there are moves towards deep learning systems which are thought to be far more accurate when making predictions (though less transparent)[11]. I am going to outline the

[10] An explanation of an artificial neural network is in §2.6.3.
[11] I introduce these models in §1.4.

statistical and deep learning methods relevant to criminal justice risk prediction in this section

2.6.1 Decision Trees and Random Forests

Decision trees are one of the oldest methods used in machine learning (Alpaydin 2016, 77). This method does not look at a decision at the end of the set of data, but rather it makes a decision for every individual step (Chen 2015, 18). At each step the algorithm performs an "if-then" function. That is to say, at each step the model will make a decision based on what has occurred at that step.

The tree is:

> [C]omposed of decision nodes and leaves; starting from the root, each decision node applies a splitting test to the input and depending on the outcome, we take one of the branches. When we get to a leaf, the search stops and we understand that we have found the most similar training instances, and we interpolate from those.
>
> (Alpaydin 2016, 78)

The best way to explain this is by way of an example. Imagine Frederick has committed an offence and a decision tree is being used to assess his risk of reoffending. First Frederick might be asked his age, if under 25 he will be directed to answer one set of questions, if over he will be asked another. Say Frederick is over 25, he might then be asked his age when he committed his first offence, say it was when he was 18, his next question might then be whether his postcode was 1030, 1091, or 3921 or whether it was any other postcode. The paths will continue to split until all questions have been answered at which

point the system will tell us Frederick's likelihood of reoffending base on the similarity of his answers to other offenders.

Whilst single decision trees are not generally used in algorithmic risk prediction, a statistical method used for predicting recidivism rates called random forest is used (this is the method used by those who operate Durham's HART recidivism prediction tool). A random forest is essentially a collection of decision trees. Random Forest models are a number of decision trees that are trained on random subsets of data whose results are then bundled together for a final output (Alpaydin 2016, 79) (Refer to Figure 1. for an example of a criminal justice random forest model). This type of model allows for a 'smoother estimate' (Alpaydin 2016, 79). The estimate is smoother in the sense that when you use a forest of non-identical trees and take an average from this you have a much more precise prediction than if you were using a single tree (Fry 2018, 58). Some advantages of this type of model over other types of model is that it can detect rare but dangerous outcomes and it does not model in a linear way, which is beneficial when modelling with multiple variables (Oswald et al. 2018, 227). Further strengths of random forest models include the ability to include a large number of predictors, taking into account forecasting errors when developing new models, and it is able to predict beyond binary outcomes, which in the criminal justice case means it can predict more than just whether someone will reoffend or not, but perhaps in what time period they are likely to offend (Barnes and Hyatt 2012, 2).

Random forest models are easier to provide an explanation for than say an artificial neural network in the sense that the 'rules are easy to interpret' (Alpaydin 2016, 78). However, this does not mean the explanatory task is trivial. Often random forest models have so many decision points that it becomes impossible to provide a complete explanation. For example, Durham's HART model has 4.2 million decision points (Oswald et al. 2018, 12).

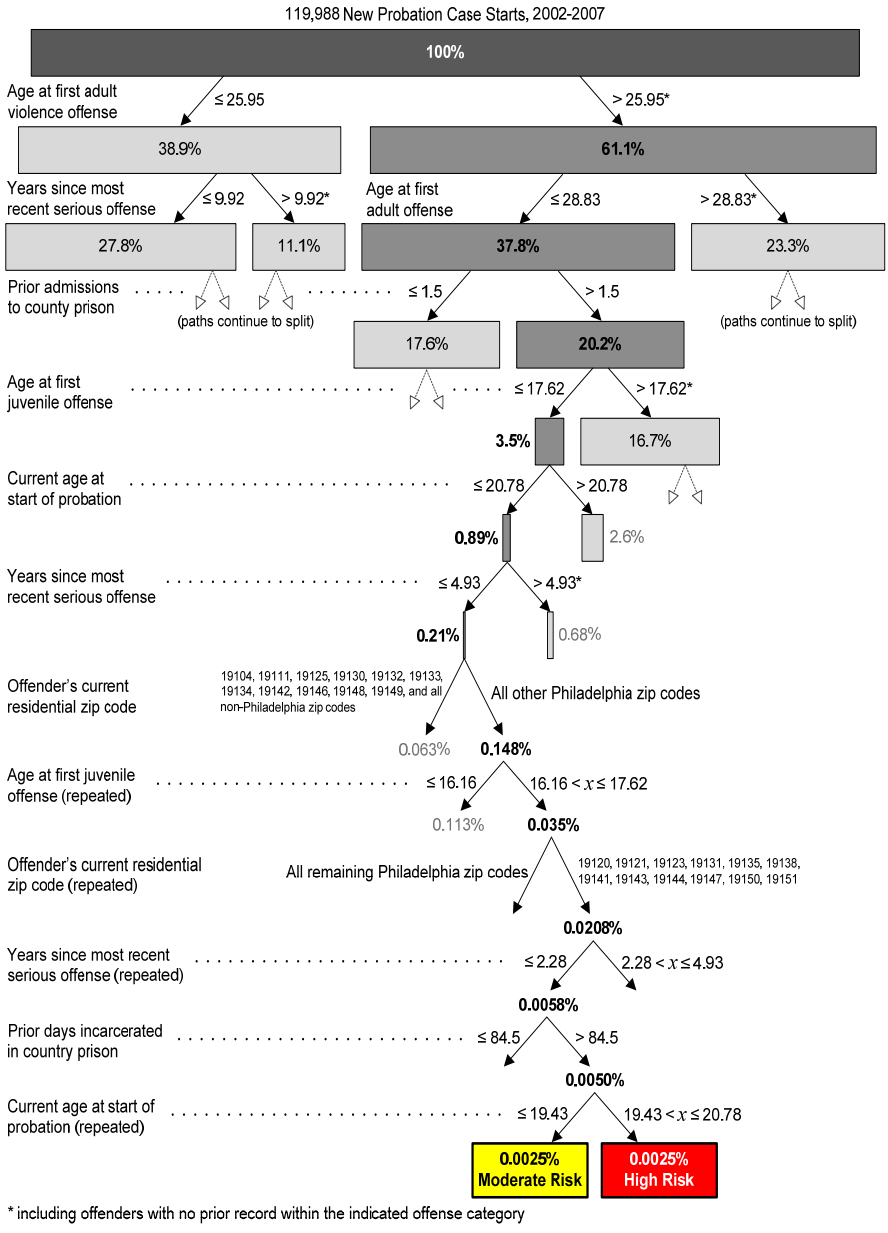

Figure 1: Criminal Justice Random Forest Model

(Barnes and Hyatt 2012, 35)

2.6.2 Regression Analysis

The second, and perhaps more common, statistical method used for developing risk of recidivism scores is called regression analysis. It is used in both the United States' COMPAS and New Zealand's RoC*RoI prediction tools (both of which I discussed in §2.4). A regression model can be understood as a model of relationships between a set of predictor variables (say in the case of criminal justice variables include: substance abuse, mood, criminal history) and an outcome (in the case of criminal justice, the likelihood of recidivism in a certain time frame).

Regression analysis is an attempt to draw a "best fit" line that describes the relationship between inputs and outputs. In regression analysis the performance of the algorithm, that is, its predictive accuracy, 'depends on how close the model predictions are to the observed values in the training data' (Alpaydin 2016, 38–39). The idea here is that the 'training data reflects sufficiently well the characteristics of the underlying task' (Alpaydin 2016, 39). RoC*RoI and COMPAS use logistic regression, which is a type of regression analysis, to make risk assessment predictions (Bakker et al. 1999, 13; Larson et al. 2016). Logistic regression is used to describe data and explain the relationship between a dependent binary variable and multiple independent variables (McDonald 2014). So, when a logistic regression model is being used to make predictions in the criminal justice system the dependent binary variable is whether or not the offender will re-offend. The multiple independent variables are those dynamic and static factors that are thought to contribute to an offender re-offending such as their criminal history, substance abuse, and antisocial behaviour.

Cox Regression is a particular form of logistic regression model used in the COMPAS risk model in many states in the United States (Brennan, Dieterich, and Ehret 2009, 34). Cox Regression models predict the association between 'survival time' (time not

offending) and one or more predictor variable (Larson et al. 2016). Essentially the model predicts, using multiple variables (determined by a questionnaire completed by the prisoner), the length of time for which a particular offender will still not have reoffended (See fig.2 below for an example).

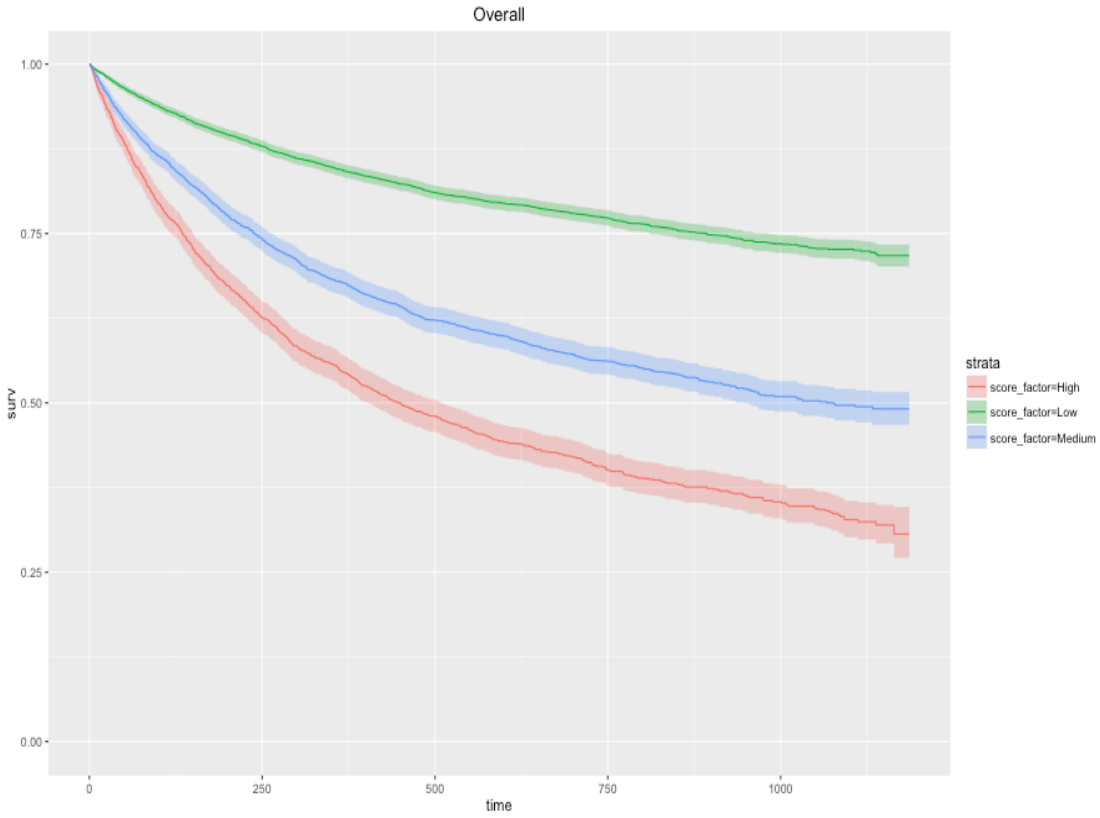

Figure 2: Cox Regression model which shows change of risk of re-offence over time (Larson et al. 2016)

Certain regression analysis methods, such as Cox Regression, appear to be more accurate predictors than a random forest model, which is to say they have a higher probability of correctly assessing an offenders risk of recidivism (Larson et al. 2016; Oswald et al. 2018, 7). However, whilst they may be more accurate, their results

(outcomes) are also not easily explainable[12]. This is because you can have such a large number of factors/variables that it becomes impossible to report them all. It seems that "accuracy" and "explainability" are in tension. There is, then, at some point, a trade-off to be made between accuracy and transparency (as I state in the previous chapter).

2.6.3 Artificial Neural Networks

Whilst I am unaware of any criminal justice predictive tools that currently use deep learning methods for predicting recidivism there are moves in this direction afoot (Pullar-Strecker 2012). The particular deep learning method I have in mind is called an artificial neural network.

The reason I find this method (whilst not currently used) worth mentioning is that explaining how a predictive risk tool using this method develops a risk score poses an interesting challenge because of its complexity. Whilst artificial neural networks are built using code, the coding does not exist in the form of mapping inputs to certain outputs. They are not like traditional algorithms that 'had their rules and weights prespecified' by a human being (Zerilli et al. 2018, 5) In fact, artificial neural networks function more like a human brain, which does not have set rules known by the human-being in question for getting from input to output. As a result, providing an explanation of how the output produced by an artificial neural network came about can be very difficult. Thus, if there are requirements for transparency of algorithms used in the criminal justice system artificial neural networks may face a big problem.

Artificial neural networks used in many domains are thought to be more accurate forms of prediction tool than random forest and regression analysis models, and are in that

[12] This is by no means stating that random forest models are easy to explain. Rather that the decision process is easier to trace. Random forest models are not "easy" to explain as there are so many decision points.

sense enticing (Tosun, Aydin, and Bilgili 2016, 3088; Teshnizi and Ayatollahi 2015, 299). An artificial neural network is inspired by the way that biological nervous systems work (Stergiou and Siganos 2011). They function much like a human brain whereby there are a number of neural pathways complete with neurons and synapses. Most commonly artificial neural networks consist of three groups or layers of what are called nodes or units. This includes an input layer of units, which is connected to one or more layers of hidden units, which is connected to a layer of output units (Stergiou and Siganos 2011) (See Fig 3. for an artificial neural network diagram).

Artificial neural networks have a 'remarkable ability to derive meaning from complicated or imprecise data [and], can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques' (Stergiou and Siganos 2011). They cannot be programmed, rather they learn by example. Because of this the examples need to be selected very carefully 'otherwise useful time is wasted, or even worse, the network might be functioning incorrectly' (Stergiou and Siganos 2011). If bad examples were picked in the criminal justice case, then it might be found that the predictions being made about the risk of re-offence were a long way off the mark.

Artificial neural networks can also learn online, 'by doing small updates on the connection weights as we see training instances one at a time' (Alpaydin 2016, 90). This is a benefit as if the 'underlying characteristics of the data change slowly [which they undoubtedly do in the criminal justice system]…online learning can adapt seamlessly, without needing to stop, collect new data, and retrain' (Alpaydin 2016, 91).

Whilst they are more accurate artificial neural networks pose a problem in terms of transparency, and in addition its 'operation can be unpredictable' and unexplainable by developers and operators (Stergiou and Siganos 2011). Artificial neural networks are so complex that providing an explanation for how a risk score was determined may well be

impossible. If these systems are complex enough that they cannot meet the ethical requirements for transparency as set out in chapter four, then artificial neural networks may not be able to meet this criterion.

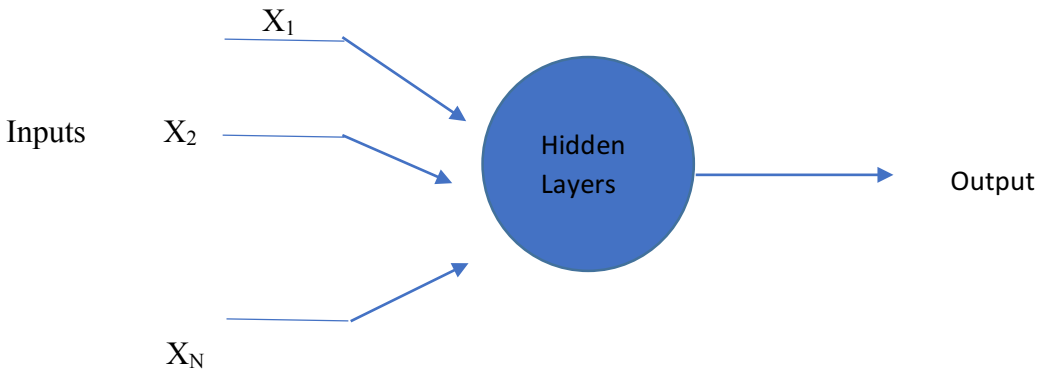

Fig.3 Simple diagram of artificial neural network

(Adapted from Stergiou and Siganos 2011).

However, as researchers such as Edwards and Veale maintain, there are pedagogical explanations of artificial networks, which is to say that it is possible to have a model-of-a-model, which can extract an explanation (Edwards and Veale 2017, 22). The idea with these models is that in order to understand how a predictive model works another model can be used to make a simpler model of the first model. Which is to say it is possible to have a statistical model that will be a simpler version of the first model which can allow the developers and operators to understand how the risk score was developed. Whether this is an appropriate way of providing an explanation, whether it will accurately explain the algorithmic model, and whether it will explain the necessary details is another question[13].

[13] I will not discuss how a model-of-a-model works in detail in this thesis. See (Edwards and Veale 2017, 22) or (Ribiero et al. 2016) for more discussion. It is an interesting question, as to whether this approach would work, and if it worked whether it would be appropriate for transparency requirements or not.

2.7. Conclusion

In this chapter I have focused on explaining how criminal justice algorithms work by describing the various generations of risk assessment tools, outlining static and dynamic factors, and describing the types of models that systems such as HART, COMPAS and RoC*RoI use, as well as the sort of models that may be used for recidivism risk prediction in the future. The purpose here is to further develop our understanding of criminal algorithms such that it gives us an indication of why the outputs of these algorithms are technically difficult to explain. In the following chapter I will focus on reasons for transparency, which is to say, arguments for why transparency is instrumentally usefully for operating criminal justice algorithms well.

3.0. Chapter Three: Arguments for Transparency

3.1. Introduction

In section 1.6. I introduced several reasons why transparency might be important when it comes to operating algorithms well. These reasons included public trust, accountability, appealing decisions, and informed consent. The question is whether these reasons do in fact bring about a requirement for making elements of an algorithm or its operation transparent. The primary purpose of this chapter is to look at these reasons in detail. I will assess whether these reasons are ethically or legally important. I will then explain how transparency applies to each of these reasons. At the end of each of these sections I will make a recommendation as to what level of transparency is required.

3.2. Public Trust

3.2.1 Introduction

It is important that citizens trust the government and its agencies and a way to increase trust is to be transparent about government processes. It is for this reason that criminal justice agencies should consider making aspects of the criminal justice algorithms that they use work available to citizens. In this section I will first argue that a democracy requires trust. I will then claim that keeping citizens "in the dark" about how an algorithm works and is used will reduce trust in the government agency operating a criminal justice algorithm

3.2.2 Democracy and Trust

For a democracy to function we need at least the majority of citizens to trust the system and transparency helps to achieve this (Eubanks 2018, 184). Without some level of trust there is a risk of citizens withdrawing support. If citizens do not trust their government unease is created between the government and the citizens, which compromises the representative

democracy (van der Meer 2018, 1) This means that a government, which is meant to be a representative of the people, can becomes weak and loses power. As Zerilli et al. state, often knowing even a rough set of reasons for a particular decision will 'engender trust in the process that led to it' (Zerilli et al. 2018, 5). Van der Meer identifies four elements of trust that exist between citizens and their government:

> (a) trust in the object's [government's] competence to act in the subject's [citizen's] interest, (b) trust that the object is benign to the subject, (c) trust that the commitment of the object can be enforced by the subject or that the object can otherwise be held accountable, and (d) trust that the behavior of the object is predictable
>
> (van der Meer 2018, 5)

Consider the use of an opaque criminal justice algorithm. It is easy to see how the use of such an algorithm by a government agency might threaten these important elements of trust. Without knowing what the algorithm is using to produce an output, nor how accurate it is, a citizen might doubt that the government is acting in the citizen's best interest. Data subjects, when subject to the algorithm, might also doubt that the government is looking out for them and acting reasonably. For example, a person with an unexplained high risk of re-offence score is unlikely to think that the government is acting fairly. Citizens also might doubt that someone will be held accountable for the outputs an algorithm develops. This is particularly concerning if the algorithm is producing several false negatives or positives (this will be discussed in more detail in section 3.3). Finally, a citizen might doubt that the government is behaving in a predictable manner when it comes to the operation of an opaque criminal justice algorithm and if the government is perceived as unpredictable then public

trust is compromised. Note that, although the government might well be acting predictably, if the algorithm and how it works is not visible to citizens, they are unlikely to understand that it is used in a predictable fashion. In order to maintain some level of trust some kind of transparency is clearly important. However, it is important to look at limits on transparency before recommendations are made.

3.2.3 Concluding remarks

Public trust is an important element of a functioning democratic government. In this section I have argued that in order to maintain public trust details need to be shared with the public about how the algorithm works, and how it is used. Although in this section I recommend that some level of transparency is important I have not suggested what that should be. This will be revisited in chapter four.

3.3. Accountability

3.3.1 Introduction

Governments and private entities must be held accountable for their actions in order to ensure good behaviour and transparency can help to achieve this. As Thomas Bivins so clearly states, humans expect and search for, accountability (Bivins 2006, 19). It is unsurprising then that people expect that someone should be accountable when it comes to the operation of predictive risk algorithms. To ensure accuracy, decreased bias, and good behaviour of developers and operators elements of how an algorithm works and how it is used should be made available to the public.

In this section I will explain what it is to be accountable. I will argue that accountability is necessary when it comes to government and private entities, and that criminal justice algorithms fall within this category (and therefore someone or some group

must be accountable for their outputs). I will then argue that making certain details of an algorithm available to the public will be a way to help ensure that those who develop and operate algorithms remain accountable and are motivated to act well.

3.3.2 What is it to be accountable? How do people become accountable?

A common claim in artificial intelligence literature is that accountability is important as it can both function both as a control for behaviour and as a means to satisfy the public that there is some group that is liable for an algorithmic system (EPFL IRGC 2018, 15; New and Castro 2018, 1; Mittelstadt 2016, 10). The questions are, what is accountability? How is it different to responsibility? And how does someone become accountable in a professional context? When a person or group is accountable for something they are liable and answerable for that thing (Hood 2010, 989). Accountability is external insofar as the "account" must be to someone separate to whoever is being held accountable (Mulgan 2000, 555). Mulgan claims that accountability also requires social interaction in that 'one side, that calling for the account, seeks answers and rectification while the other side, that being held accountable, responds and accept sanctions' (Mulgan 2000, 555).

People commonly ground accountability in responsibility, or use the terms interchangeably (Barker 2001 ,138). This creates confusion when explaining how a person becomes accountable. The following example demonstrates the difference between responsibility and accountability, and that accountability can exist without responsibility. Rory has a dog. One day Rory and his dog are walking through a park, when his dog unpredictably lunges at a passer-by and bites them. Rory is not responsible for the dog's action, after all he never saw it coming, and he did not do it. However, it is fair to claim that Rory is accountable for the dog's actions, as when he signed up to having a dog he knew it was a possibility that the dog could act out.

So, what is it that makes a person accountable? A person or group can become accountable in several ways, including by contract, through having a certain relationship, or through holding a certain role. In order to be accountable in a professional context, the person held to account for an action that has caused someone harm must a) be aware of the role that they hold and that the role requires being accountable for certain things and, b) be holding the role without being forced and with knowledge of what the role entails.

3.3.3 The value of accountability

Accountability is an important aspect of governance and business. It is an effective control for maintaining ethical standards. One argument for the importance of accountability in governance and business is that it discourages bad behaviour through control (Mulgan 2000, 556). If a person or a group is accountable for a certain action they are less likely to veer away from what is expected of them as they fear being reprimanded for bad behaviour. Essentially, as Wanna states, it 'impress[es] a self-imposed discipline on governments' (Wanna 2018, 12). A second argument for accountability is that it can promote several desired things, such as 'public disclosure, public insight into decision-making, scrutiny and evaluation, and democratic oversight' (Wanna 2018, 12). The third, and final argument, is that accountability is essential because when a data subject has been wronged there is a person or group identified as having to 'respond and accept sanctions' (Mulgan 2000, 555).

3.3.4 Accountability in criminal justice algorithms

I have argued that, in government and private corporations, someone, or some group, must be held accountable for any harms that might occur. Criminal justice algorithms are used by government justice departments and their use has the potential to cause harm, as is the case

when a criminal justice algorithm produces an output that turn out to be a false positive. The harm in this example occurs if the false positive influences a judge to sentence more harshly or a parole board to not grant parole. Having someone, or some group, held accountable for the operation and output of an algorithm will mean that a harmed data subject or their representative can receive the relevant explanation if they have appealed the use or accuracy of an algorithm[14]. This is because the overarching administrator can identify who is accountable for the harm. Moreover, if developers, operators, and policy makers are held accountable for an algorithm they will be motivated to act well when developing and operating it. This is because, if they need to be able to explain or justify judgements or actions to an affected group (or individual) they will want to try to reduce harms, and act honourably.

3.3.5 Is transparency necessary for ensuring accountability in criminal justice algorithms?

Transparency is necessary for keeping a person or a group accountable and behaving as they should (Zarsky 2013, 311). If there is no transparency a deviant, or lazy, developer or operator could lie to a data subject or try to get away with things that they should not. For example, not taking steps to reduce bias or keep the algorithm updated. By making the system more transparent data subjects can see that things are running well and in addition the operator and the developer are motivated to behave well and update the algorithm because their work and actions are 'on show'. Transparency in this sense can almost act as a form of 'surveillance' (Heald 2006b, 28).

Some theorists, such as Kroll and Hood, have argued that transparency is not a necessary component of accountability (Kroll et al. 2017, 665; Hood 2010, 991). Kroll et al. claim that the necessity of transparency for keeping people accountable can be avoided when using algorithms by having what he calls cryptographic commitments (Kroll et al. 2017,

[14] By representative, I mean a lawyer, or similar.

662). Cryptographic commitments are 'a digital equivalent of a sealed document held by a third party or in a safe place' (Kroll et al. 2017, 665). Essentially the developer of the algorithm can give the source code (or similar) to a third party. The third party then holds on to this information and after an agreed time period it can be revealed to the data subject to show the information that was used (Kroll et al. 2017, 665). The proposed benefit of this is that if the commitment is made then the algorithm developer and operator must stick to it. Knowing that this information can be accessed later should motivate the operator and developer to act in a way that is honest, unbiased, and as accurate as possible when operating and developing an algorithm. However, this does not mean that transparency is avoided, as Kroll claims (Kroll et al. 2017, 662). A cryptographic commitment is simply a different form of transparency. It still requires that there is information made available to those who are affected by it at some point. It becomes an *ex post* form of transparency where the data subject receives information about the system post decision.

A point against using cryptographic commitments that should outlaw their use in a criminal justice context is that they leave open the possibility that there is a harmful action to a data subject now and the public will not discover it until the information contained in the cryptographic commitment is made available. This would mean that the developers and operators could not be challenged and held accountable until that point which may be far beyond the point that is useful to the data subject. Furthermore, those accountable might not be there to hold to account later.

3.3.6 Details the public should know in order to motivate those who are accountable

Having demonstrated the need for accountability and that transparency is essential to motivating those accountable, I will make a couple of transparency recommendations. There are a several pieces of information that must be made available as a way of motivating those

accountable. Firstly, the criminal justice algorithm's developers and operators need to establish some reasonable limit to how many false positives and false negatives can be accepted. That the algorithm is meeting these requirements should be information that is made available to the data subjects (even if the numbers are not made available to the public).

It must also be made clear to the public that an unfortunate consequence of predictive risk of re-offence algorithms is that there is a missing a section of data. The number of false positives, that is, those who are said to be high risk and yet would never reoffend, is unknown. It is important to make clear that this section of data is missing when presenting data on the accuracy of algorithms.

Secondly, information must be available regarding the methods that are in place to ensure that the algorithm is as unbiased, accurate, and updated as possible. This may appear in form of an audit on what measures have been taken to reduce bias, increase accuracy, and update the system.

In cases where a decision cannot be made about who is accountable, the algorithm should not be in operation. The safeguard of accountability must exist before a criminal justice algorithm is utilised. Also, as O'Neill also aptly points out, people often assert that transparency is achieved simply 'by placing reams of documents on websites, so making them 'available' to the public at large' (O'Neill 2004, 270). This is not an appropriate way to approach transparency for accountability as many members of the public lack the capacity to make sense of information presented in this way (O'Neill 2004, 270). Instead, this information needs to be expressed as simply as possible with explanatory notes where it is thought that confusion might arise. Making this information available can help to reassure a data subject that things are done well or in an acceptable way and also ensures operators and developers act well.

3.3.7. Concluding remarks

This section has shown that accountability is important in the context of operating and developing criminal justice algorithms. It has also shown that transparency is essential for motivating those accountable and that certain aspects of the algorithm need to be transparent in order to maintain accountability. Transparency and accountability have the resulting effect of ensuring accuracy, fairness, and that the algorithm rests on quality data. This section concludes by outlining recommendations for what what elements need to be transparent in order to ensure that those accountable are motivated to operate the algorithm well.

3.4. Rights to appeal and review

3.4.1 Introduction

If you think something has gone wrong regarding a decision made about you in the justice system there are safeguards in place to help you check that the process was in fact fair and that your rights were properly observed. In New Zealand, and in fact in many common-law countries, there are at least three ways (depending on the context) in which a decision might be questioned. These include a right of appeal, judicial review, and statutory rights such as 'Review of decisions' in section 67 of New Zealand's Parole Act 2002 (*Parole Act* 2002, sec. 67). These legal rights protect citizens by allowing them to challenge decisions that are made about them. However, an appeal or request for review requires a certain level of transparency about the decisions that were made in order for an offender to challenge a decision (Kehl et al. 2017, 28). In this section I will explain these different rights and their relevance to the use of predictive risk algorithms that are used in criminal justice systems. I will also look at two appeals cases that have involved the use of algorithms, one in the United States and one in New Zealand. I will conclude this section by using the State v Loomis

[2017] and R v Peta [2007] appeals cases to argue that transparency is important for citizens to be able to properly utilise these rights.

3.4.2 A right of appeal

A right of appeal is a feature in most if not all common-law countries. An appellant or defendant in a case that comes in front of a court has a right to appeal a decision that is made about them. This appeal can be based either on an error in the facts that are used or a question of law (Rodriguez Ferrere 2012, 829). An error of facts occurs when a judge has either used incorrect facts or has interpreted facts incorrectly to make a judgment. An error of law, on the other hand, is when the judge has used the correct facts but the individual affected by the decision thinks that the judge has not applied the law in the right way (Rodriguez Ferrere 2012, 830). An appeal must then be allowed (at least in New Zealand law) if '(a) for any reason there is an error in the sentence imposed on conviction; or (b) a different sentence should be imposed' (*Criminal Procedure Act* 2011, sec. 250(2)).

A right of appeal might be utilised by an offender who believes that an error was made when a risk score developed by an algorithm was used in a sentencing decision. This could either be because they think that the algorithm was inaccurate or because they think that using the algorithm meant certain rights the individual holds were not observed by a judge. An example of a right of appeal based on the use of an algorithm is the *State v Loomis* case in which Loomis appealed the decision made about him because he believed that the court, in having used a risk score produced by the COMPAS algorithm, had not properly observed his right to due process.[15]

[15] Due process can be understood as a legal proceeding that protects an individual's rights (Bird 1913, 45).

3.4.3 Judicial Review

Judicial reviews are a way of reviewing decisions that are made by public entities such as government departments or parole boards (*Judicial Review Procedure Act* 2016). These decisions, at least in New Zealand, are reviewed in the High Court (*Judicial Review Procedure Act* 2016, sec.8). Judicial reviews are a way to challenge the procedure used to make a decision rather than to look at whether the outcome of the decision was right or wrong (Public Law Project 2006, 1). The High Court will not substitute a 'new decision', instead they just look at whether the way in which the decision was made was appropriate. After reviewing the decision the High Court can choose to quash the agency's decision, order the agency to undertake a particular action, order a new decision or order an injunction (*Judicial Review Procedure Act* 2016, sec.16). Decisions that could come under a judicial review could for example, be an immigration decision, or about a prisoner's rights. In the case of algorithm use, if an offender thought that a parole board had not followed the right procedure when using the risk score produced by the algorithm, they could request a judicial review. However, this is unlikely to be a common approach given that, at least in the New Zealand case, an offender has a statutory right for reviewing a parole board decision.

3.4.4 Statutory Rights

Statutory rights are rights that are designed to protect the individual through statute and are granted by a local or national government. One use of statutory rights is to give an individual a right to review a certain type of decision made about them. This gives the citizen the chance to challenge a decision if they think that they have not been treated fairly. One example of a statutory right that an offender would use to challenge a decision is Section 67(3) of New Zealand's Parole Act 2002 (*Parole Act* 2002, sec. 67(3)). An offender is given the right to

have the parole board review a decision they have made about the offender. They can ask for the review if they believe that the parole board in making the decision:

a) Failed to comply with the procedures set out in [the Parole] Act and any regulations made under it; or
b) made an error of law; or
c) failed to comply with a policy of the Board developed under section 109(2)(a), which resulted in unfairness to the offender; or
d) based its decision on erroneous or irrelevant information that was material to the decision reached; or
e) acted without jurisdiction

(*Parole Act* 2002, sec. 67(3))

Essentially, if an offender had reason to believe that the use of an algorithm had led to any of these conditions being met they would have grounds to ask for a review based on the use of the algorithm. Again, this review does not lead to a new decision being made; rather it looks at whether the decision making fairly lead to the outcome. It is only when this right to review has been exercised, the parole board has reviewed their process, and the offender still believes that the outcome was not right, that an appeal can be made to the High Court in regards to a parole board decision.

3.4.5 An appeals case: Loomis vs. Wisconsin Supreme Court

Appeals based on algorithms used in sentencing decisions have already occurred. In 2013 Eric Loomis made an appeal to the Wisconsin Supreme Court in relation to the use of an algorithm. Loomis claimed that the 'COMPAS assessment infringed on both his right to an

individualized sentence and his right to be sentenced on accurate information. Loomis additionally argued on due process grounds that the court unconstitutionally considered gender at sentencing by relying on a risk assessment that took gender into account' ("State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing" 2017).

Justice Ann Walsh Bradley rejected Loomis' claims that his due process rights had not been appropriately observed by using the COMPAS algorithm in his sentencing decision. She gave a number of reasons as to why she had rejected Loomis' appeal. Firstly, Bradley found that 'gender as a factor in the risk assessment served the non-discriminatory purpose of promoting accuracy' ("State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing" 2017). Bradley also stated that 'COMPAS uses only publicly available data and data provided by the defendant, [and] the court concluded that Loomis could have denied or explained any information that went into making the report and therefore could have verified the accuracy of the information used in sentencing' ("State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing" 2017). What Bradley is essentially saying here is that Loomis knew the information he had provided, and the other information was publicly available, so he should have known whether the information that the algorithm used was accurate. Bradley however, did make a point of stressing the importance of a defendant experiencing individualised sentencing and did admit that COMPAS does only make a decision based on aggregate data on risk for 'groups similar to the offender' ("State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing" 2017).

The responses given by Bradley are problematic for a few reasons. Firstly, to state that 'Loomis could have denied or explained any information that went into making the

report and therefore could have verified the accuracy of the information used in sentencing' is missing the point of Loomis' claim ("State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing" 2017). Loomis makes his appeal partly on the basis that he does not know the accuracy and scientific validity of the algorithm. Yes, Loomis did know the information he was providing, and could access the other information they had used about him, but this does not actually lead to Loomis knowing whether the prediction made by the algorithm was made on accurate bases. Given that the inner workings of the algorithm are not made available, Loomis has no knowledge of what answers were used in the prediction, nor the weighting each answer had. This means that Loomis has very little understanding of whether the risk score produced was accurate or not.

Secondly, the question arises about what safeguards are in place to ensure that a judge or parole board does not place too much emphasis on the score, and that the sentence is sufficiently individualised. A judge may well, either unconsciously or consciously, place more weighting on the score than is acceptable. Bradley stated that risk scores cannot be used 'to determine whether an offender is incarcerated' or 'to determine the severity of the sentence' ("State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing" 2017). However, a judge might unconsciously, or consciously do so.

3.4.6 A New Zealand Case: R V Peta [2007]

New Zealand has also had appeals cases in relation to algorithm use, but only about decisions made by the parole board as algorithms are not used in sentencing decisions in New Zealand. One case was *R. v Peta* (*R v Peta* [2007] NZCA). However, instead of the offender (Peta) making an appeal based on the use of the algorithm, the appeal *found* that the risk scores

produced by the algorithms had not been used properly, which had led to an unfair outcome for Peta. To give some context, Peta had been given an extended supervision order (ESO) due to previously 'committing an indecent act with a girl under 12' (*R V Peta* [2007], NZCA sec. 1). Due to being perceived as high risk he was given a 10 year ESO. Peta appealed the length of this ESO and following this it was discovered that the reasons that had been given for the ESO were far too inadequate. Some of these 'reasons' were the risk scores that various algorithms had produced.

There were a couple of algorithms that were used to assess Peta's likelihood of re-offfending: Sex Offender Needs Assessment Rating (SONAR) which assesses using dynamic factors, and Static-AS or Automated Sexual Recidivism Scale (ASRS) which assesses, as the name indicates, using static factors (*R v Peta* [2007] NZCA). In the appeal it was discovered that the original assessor had claimed to use SONAR factors, but had not actually done so. Rather the original assessor had used Stable 2000 factors, which is a risk of sexual recidivism tool used in Canada (Hanson et al. 2007, i; R v Peta [2007] NZCA, sec. 66). According to Riley, the 'director of the Psychological Service of the Department of Corrections' who was called by the Crown, this would not 'necessarily have invalidated his findings'. However, he stated that the assessment was not properly conducted or scored (*R v Peta* [2007] NZCA, 66). Furthermore, it was found that Peta's ASRS score indicated that he was at higher risk of re-offense than those processing the appeal thought he was. When looking at other clinical judgements and assessments based on dynamic factors Peta appeared far less high risk. Vess, a 'senior lecturer in abnormal psychology and adult mental health at Victoria University', called by Peta, claimed that there were also missing factors in the ASRS such as whether there had been previous sex offences, or whether the victim of the offence was a boy (*R v Peta* [2007] NZCA, sec. 15,80). They admitted that Peta is at high risk of re-offense if he does not deal with his substance abuse but they state that that his risk of sexual

reoffence is low (*R v Peta* [2007] NZCA, sec 91). As a result, due to the perceived way in which the algorithms were used by the parole board, the appeal was allowed, and 'the ESO imposed on Mr Peta [was] quashed' (*R v Peta* [2007] NZCA, sec. 105).

3.4.7 Why transparency is important in cases of appeal and review

To this point we have seen that appeals can be made regarding the use of criminal justice algorithms. We may now turn to the question of the importance of transparency in cases of appeal and review. A certain degree of transparency is important so that a defendant has the information available to exercise their right of appeal or request a review based on the use of an algorithm (Zerilli et al. 2018, 4). If there is a fault in an algorithmic system's accuracy or the algorithm's use is thought to conflict with a defendant's rights, then there are grounds for an appeal. However, if the defendant has no access to some of the details of the algorithmic system they would not have appropriate evidence to form an appeal. Given that a person can appeal the evidence used in a decision, they might need to know details about the algorithmic system, such as its accuracy, what factors it uses, and how the risk score is developed by the algorithm. Furthermore, the decision maker should be as transparent as possible with regard to how they have used the risk score in their decision making (although I do recognise that humans cannot always identify how something influences their decision making).

Additionally, as part of the appeals process there must be a judgment given with reasons for the outcome of the appeal (*Criminal Procedures Act* 2011, sec. 340-342). This explanation will also require certain kind of transparency. The *R v. Peta* case is a good example of the type of explanation, and level of transparency, that should be expected in an appeal outcome, whereas the Loomis case shows where transparency is lacking. In the Loomis case, Loomis was unable to access 'information related to how COMPAS weighed

particular input variables and how these inputs were calculated for his final risk score, because the developer considered such information a trade secret' (T. R. Moore 2017, 17). Here, the court justified the lack of transparency on the grounds of commercial value. This lack of transparency based on intellectual property rights brings about a tension between satisfying a person's right to an appeal and a company having trade secrets (this will be discussed in §4.2). In contrast to *State v. Loomis* [2017], the *R v Peta* [2007] appeals case is extremely open. The methods for the risk assessment were made openly available in the appeal and the reasons for why the assessments were problematic in this case were openly discussed and explained in a clear manner (as I have demonstrated in §3.4.6. section). This was an appropriate approach to explaining why the appeal went ahead.

3.4.8 Concluding remarks

In this section I have shown that an offender can request an appeal or review of a sentencing or parole decision in New Zealand. I have also shown that the use of an algorithm could be grounds for an appeal or review if the offender believes that the algorithm is inaccurate, or the use of the algorithm has led to a miscarriage of the law or has infringed on an offender's rights in some way. I have further argued that there needs to be transparency surrounding how the algorithm works and its use in order for the offender have enough information to request an appeal or review. Furthermore, with the use of two case studies I have also shown that a report on whether the appeal is accepted or not requires a certain amount of transparency about the algorithm and how it was used.

3.5. Informed Consent

3.5.1 Introduction

Informed consent is generally discussed in relation to medical procedures, examinations, and research because of the harm that they may cause a patient. The notion of consent is also used in the 'tech world', for example, when I sign up to Facebook I must give consent for them to use some information about me. The question is whether we should require informed consent in the criminal justice algorithm case. In this section I will introduce the notion of informed consent and discuss three ways in which consent is often justified, these being: autonomy, protection from harm, and trust. I will then argue that there are some cases in which data subjects might be harmed and that, in such instances, we might need to require consent. Subsequently, I will argue that, in cases where we do need informed consent, transparency will be required in order to make sure the consent is truly informed.

3.5.2 What is informed consent?

Informed consent does not need much explanation as a concept as the name is essentially self-explanatory. Informed consent can be best understood as the act of a subject giving permission to an actor (generally in a professional role) to do something to them that may harm them. For the consent to be informed the subject must have relevant information available to them, which includes the process as well as the risks, and benefits of supplying information or permitting an action. This allows the subject to decide what is right for them. Scholars often justify the necessity for informed consent based on three concepts: protection from harm, autonomy and trust (Eyal 2019).

a) Protection from harm

The primary justification for having informed consent is that it can act as a form of protection from harm for a subject (Tadros 2011, 23). A subject can be harmed in a

number of ways. For example, an action itself can cause physical harm to the subject, and a subject can be taken advantage of for the progression of science, or an administrator's career (Eyal 2019). Requiring informed consent can act as a way to protect the subject from being coerced or unknowingly signing up to be harmed (O'Neill 2003, 5) It can also allow them to acknowledge the harm or potential harm, before going ahead with an action.

b) Autonomy

Another way to justify informed consent is through autonomy (Eyal 2019). Autonomy is the idea that adult humans who are sufficiently cognitively aware are self-aware agents and should be able to govern themselves (Downie and Telfer 1971, 301). Informed consent can then act as a means of governing oneself. One example which is often used to demonstrate what informed consent can do for exercising autonomy, is that of a Jehovah's Witness requiring a blood transfusion to survive. Jehovah's Witnesses will not accept another person's blood products even if their lives depend on it as doing so conflicts with their religious beliefs (Eyal 2019). By requiring that the patient give consent prior to the procedure the Jehovah's Witness can decline the treatment. To not give the patient a blood transfusion is not the option that a doctor would take if consent were not required. However, an emphasis on the value of autonomy suggests that it is important that the patient has been able to follow what is important to them even if by doing so they will be physically harmed.

c) Trust

The final often-cited way for justifying informed consent is that it can increase trust in a subject (Eyal 2019). This is particularly relevant in the medical procedure case. Giving a person the option to make a decision for themselves may increase their trust in the system as a whole (Tännsjö 2014, 445). For example, if a person is given the

choice to opt out of a procedure they may feel less forced by a medical professional. This may mean that in the future the patient continues to visit the doctor because they know that they will not be forced into things that they do not want to do.

3.5.3 Informed consent and algorithms

The question is whether any of these justifications for requiring consent fit the criminal justice algorithm case. I will take each of these justifications and discuss whether the algorithm case fits.

a) Protection from harm

An offender could be harmed by the use of a criminal justice algorithm. The most obvious of harms is that an offender could receive a risk score that completely fails to reflect their level of risk due to the algorithm for some reason producing a very inaccurate risk score. A decision maker could then be influenced to give the offender a harsher sentence or not give them parole based on the risk score. One might object to this, by arguing that a judge or parole board will have other information to make a decision so the algorithm will not necessarily do harm. However, judges and parole boards may be increasingly less likely (as algorithms become more accurate) to override an algorithm risk score, for fear that if they did not take it into account, they might release an offender who goes on to, for example, to be a mass murderer.

b) Autonomy

In the criminal justice case we might say that an offender should be able to exercise their autonomy and choose not to be subject to a risk score produced by an algorithm. Some might disagree with this, stating that by having offended and being a danger to society the offender has lost their right to choose to not have a risk score produced about them. However, this does not seem to be how we deal with all cases of offence

so it does not seem that this objection should hold here. Take the following example to demonstrate this. Imagine that a driver has had too much to drink and is driving home. He is pulled over by police and asked to participate in a breath test. Under New Zealand law he can choose not to consent to the side-of-the-road breath test. If a police officer has reasons for concern she can then send the driver for further testing (New Zealand Transport Agency 2014). In this case, whilst the drink driver cannot avoid testing as a whole he has an opportunity not to consent to a certain form of testing. An analogy could be drawn in the algorithm case. Whilst a risk assessment will be done in some form or another the offender could exercise autonomy and ask for an alternative form of risk assessment.

c) Trust

Trust seems like a less convincing justification for informed consent in the criminal justice algorithm case. An offender is not likely to be more compliant in the future simply because they were not forced into having an algorithm assess the degree of risk they were at of reoffending. Furthermore, a risk assessment can happen with or without the use of an algorithm, so it is even less likely that they will behave better in the future simply as a result of opting out of having the algorithm determine a level of risk versus something else determining a level of risk.

At this point the potential for harm and the ability to exercise some degree of autonomy appear to be good reasons to require that a data subject give consent for a risk score to be determined. However, as Johnston, Dare, and Gambrill point out, there are some algorithms that use only information that is available legally and without interaction with the data subject (Johnston 2018; Dare and Gambrill. 2017). For example, RoC*RoI, used in New Zealand, only uses static factors that are acquired without input from an offender (Johnston

2018). In a case like this the importance of consent is not compelling. However, in an algorithm such as COMPAS, where the algorithm relies on data that is acquired from offenders themselves, the case for consent becomes more convincing. It becomes even more compelling when we compare it to a psychological report that is also used to help assess risk. A psychologist:

> provide[s] assessment reports to the court, the New Zealand Parole Board, prisons, and probation… [t]he offender gives their consent for information to be collected from themselves and knows at the start of the process that this information may end up in a psychological report that is provided to one of the above entities
>
> (Freeman-Brown 2013, 18)

This means that the offender can choose not to give information if they would not like to, or thinks that doing so might incriminate them. Criminal justice algorithms that require information from the offender should be treated analogously. So, an offender should not have to provide the information asked for in the risk assessment.

3.5.4 Do algorithms require consent?

After a thorough assessment of the importance of consenting to calculate a risk score, the conclusion about consent requirements is far clearer. As Dare and Gambrill suggest, in cases where access to the data used is already allowed 'then it seems legitimate to regard the use of the tool at that point as a new and effective way of doing something already permitted' (on the proviso that the tool is more accurate) (Dare and Gambrill 2017). However, in cases where algorithm inputs include factors that must be acquired by communicating with the

offender it is good to take a morally cautious approach and require informed consent. After all, a user gives consent for a website to use cookies, so it is bizarre that consent is an issue in the much higher stakes criminal justice algorithm case.

3.5.5 How is transparency relevant to giving informed consent?

To achieve informed consent a certain degree of *ex ante* transparency is required, such that the subject is aware of 'all the relevant information' (Tadros 2011, 25). In the medical case the information required for consent can include information about what happens in a procedure as well as the risk and benefits of the procedure (Eyal 2019). This allows the patient to weigh up relevant factors before giving consent for the procedure. A similar set of conditions should be demanded in the criminal justice algorithm case. For the offender and potential data subject to give informed consent in the criminal justice algorithm case prior to its use they must be told several things. Firstly, they should know what the algorithm is used for. Secondly, they should be told how the product of the algorithm will be used. Thirdly, they should be given a rough indication of how the algorithm works (this does not need to include weightings or factors, rather it must explain the process an algorithm takes from input to output). Fourthly, they should be presented with the risks and benefits of providing this information.

3.5.6 Concluding remarks

Informed consent is an interesting issue for the operation of criminal justice algorithms. This section has shown that in a case where the information an algorithm needs is obtained legally and independently from the data subject it looks as though informed consent is not needed. However, if the offender must give information in order for the risk score to be produced then it would be problematic not to require consent as subjects should have the opportunity

to exercise their autonomy and not incriminate themselves. In order to do this some *ex ante* transparency is necessary.

3.6. Conclusion

This chapter has assessed the justifications for some degree of transparency. I focused on the importance of public trust, accountability, right of appeal or judicial review, and finally informed consent. I have argued that each of these elements is important when it comes to operating a criminal justice algorithm ethically and lawfully. I have then gone on to argue that each of these elements requires some sort of transparency regarding the operation or construction of a criminal justice algorithm. This has all been done with the purpose of helping to develop ethical algorithm transparency requirements. However, there may well be limits to transparency. The following chapter will focus on these possible limits.

4.0. Chapter Four: Restrictions on Transparency

4.1. Introduction

As I established in chapter three there are several good reasons for making elements of how an algorithm works, or how it is operated, available to the public or data subjects. However, there are also reasons to restrict transparency. This section will consider justifications for restricting transparency, which include intellectual property, gaming the system, invoking doubt through too much transparency, and technical limits. Each of these justifications will be explained and analysed. Where tensions arise with reasons for transparency I will offer possible ways in which they can be mitigated.

4.2. Intellectual Property

4.2.1 Introduction

When you have a squeaky door you may well fix it by spraying it with WD-40, a name and product that most of us know. The inventor was on to something good and the product became ubiquitous. What most people probably do not know is that the formula for WD40 has been kept hidden from the public in a bank vault for over 50 years (Vethan Law Firm 2016). The reason for this is that the inventor does not want others taking the formula and making the same or similar products because doing so would take away the company's competitive advantage. The WD-40 formula is what is called a trade secret, information that can be withheld from the public by the company to keep a competitive edge. One way in which developers of an algorithm justify the opacity of their algorithms is on the same basis as the above example. They claim that in being transparent they will lose their intellectual property and therefore their competitive edge. In this section I will explain what intellectual

property is and what different forms of intellectual property exist, and then argue that the use of government algorithms can be regulated in such a way that the intellectual property justification does not have to be a consideration for opacity.

4.2.2 What is intellectual property and what forms does it come in?

The term "intellectual property" can be understood as the ownership of an idea (A. Moore and Himma 2014). Owning this "idea" gives the owner a selection of rights related to it (Resnik 2003, 319). These rights differ depending on the type of intellectual property.

There are several ways in which intellectual property is recognized by law in most countries. I will define the various forms of intellectual property and then focus on trade secrets, which is the form of intellectual property that are used to protect algorithms.

Firstly, there are patents. Patents are original physical creations, that is inventions, which an inventor will allow access to for a certain period of time and in doing so gain royalties. According to Hettinger these inventions must be 'novel; constitute nonobvious improvements over past inventions; and they must be useful' (Hettinger 1969, 33). The patent protects items made using specific principles for a specific purpose, but does not protect the principles themselves (Hettinger 1989, 33). An example of a patent could be the following: imagine that Katherine has developed a self-driving supermarket trolley. This is a novel invention and there is nothing like it out there on the market. Having satisfied the criteria for patenting she can apply for a patent with a governing body. If the governing body awarded the patent, then she could collect royalties from any supermarket that used the trolley. Patents often have a time limit of 20 years. However, they are one of the strongest forms of intellectual property and will secure an inventor's idea, as well as allow the inventor to reap the financial benefits of the idea (New Zealand Intellectual Property Office 2018).

Secondly, there are copyrightable ideas. These came about as a way to protect an author's ideas expressed in books and to stop other people from taking them. It also now covers particular songs, and particular computer programmes. Copyright tends to hold for the author's lifetime plus 50 years (Hettinger 1989, 33). Copyright essentially will grant the owner legal ownership to the expression of their ideas, but not the physical copies of it. For example, if Vikram Seth's novel "A Suitable Boy" has copyright, then Seth's expression of ideas in this book cannot be reproduced without Seth's permission, or without acknowledgement.

Thirdly, there are trade secrets and trademarks, both of which are used to protect the reputations and success of companies (Steidman 1962). Trade secrets are things like recipes that are kept private to give a company a competitive edge (Steidman 1962, 4; Ministry of Business, Innovation and Employment 2018). Trademarks, on the other hand, are a way to protect something 'special' about a company, so that the company is more recognisable for a certain product. Coca Cola™ is an oft used and good example here (Ministry of Business, Innovation and Employment 2018). They have trademarked their name to make sure that their drink is more recognisable. The recipe for Coca Cola™ has also been kept as a trade secret to help protect Coca Cola's point of difference and profits. Trade secrets are an unregistered form of intellectual property under New Zealand law (Ministry of Business, Innovation and Employment 2018). This means that they are not registered with an official agency but are still protected by law. Trade secrets are protected simply by means of the secret not being shared. Anyone who finds out a formula or recipe through reverse engineering has technically not broken any rules. On the other hand registered intellectual property is lodged with a governing body (Ministry of Business, Innovation and Employment 2018). Trademarks can be registered or unregistered. A registered trademark has an ®,

whereas an unregistered trademark has a ™ (Ministry of Business, Innovation and Employment 2018).

4.2.3 Criminal justice algorithms and intellectual property

The type of intellectual property relevant when it comes to protecting algorithms is trade secrets. This is because an algorithm is abstract, so it cannot be secured with a patent[16]. Trade secrets are therefore the only form of intellectual property that can give a developer 'meaningful protection' (T. R. Moore 2017, 6). In New Zealand, a trade secret is defined as:

a) is, or has the potential to be, used industrially or commercially; and
b) is not generally available in industrial or commercial use; and
c) has economic value or potential economic value to the possessor of the information; and
d) is the subject of all reasonable efforts to preserve its secrecy.

(*Crimes Act* 1961, sec. 230(2))

To protect their competitive advantage algorithm developers often claim that their source code and other inner workins is a trade secret which means that they do not have to share it (T. R. Moore 2017,4). Northpointe, for example continues to 'keep the insides of its algorithm [COMPAS] a closely guarded secret, to protect the firm's intellectual property' (Fry 2018, 69).

[16] A patent has more stringent legal protection than a trade secret, but it requires that the object physically exists.

However, as Wexler states, ‘the introduction of intellectual property into the criminal justice system raises under-theorized tensions between life, liberty and property interests’ (Wexler 2018, 8). The trade secret justification for opacity can come into direct tension with other considerations. In essence, intellectual property claims can come in to conflict with important rights. One is that citizens may have the right to more information about the algorithm than a trade secret restriction would allow. The other problem that can arise is that a person who requires an explanation due to perceived discriminatory or inaccurate decisions may be unable to access it due to trade law protections (T. R. Moore 2017, 3). If it is important to be able to receive an explanation and there are trade law restrictions, then there is a problem.

The Loomis vs. Wisconsin case is a perfect case for demonstrating the harm that comes from information being withheld. Loomis was arrested, tried and sentenced and a COMPAS score helped to guide the sentencing decision. Loomis appealed the decision made and demanded some more information on how the algorithm came to calculate its scores. Northpointe (now called Equivant) stated that they could not give away these details because they were trade secrets. This is a problem and causes tension. Loomis needed to know some important information and the court was unable to share it because the private developer who sold the algorithm would not provide the information (“State v. Loomis” 2017)

It looks as though allowing criminal justice algorithms to be protected by trade secrets can cause problems for a data subject and for an operator of an algorithm in several ways. It does so by limiting the ability to be transparent about their use such that an offender cannot exercise rights to appeal. It also limits the operator’s ability to explain a decision which has an immediate effect on the data subject in many ways. This should motivate us to look for alternative approaches to the development of predictive risk algorithms in criminal justice contexts.

4.2.4 How to avoid trade secrets from inhibiting transparency

At this point, there looks to be a tension between justifications for transparency such as informed consent and a right to appeal and the trade secret opacity justification. The question then becomes: which set of justifications should take priority?

In order to avoid the intellectual property claim for opacity altogether government agencies could be required to develop algorithms in-house or by contract. By doing this the operators are not left at the mercy of a private developer who can choose to withhold information that will give them some sort of competitive advantage. This approach also has a further benefit: the algorithm ends up being developed for the criminal justice system it will be used in. This means that it will be developed with a specific purpose and context in mind.

One possible objection to this approach is that it reduces competition between developers. The reason that this is perceived to be a problem is that without competition people are less motivated to keep improving the algorithm. However, there are other pressures in the criminal justice system that could be put in place which would require an algorithm be kept as accurate and up to date as possible, for example, a mandated schedule for review. This would mean there would be constant motivation to develop and to improve an algorithm.

4.2.5 Concluding remarks

In this section I have explained various forms of intellectual property and indicated that the inner workings of an algorithm often fall into the category of "trade secret". I have shown that to claim that the source code is a trade secret can inhibit a data subject's ability to appeal a decision, amongst other things, as well as inhibit an algorithm operator's ability to explain how an algorithm developed a risk score about an offender. The Loomis v Wisconsin case

discussed in §3.4.5. demonstrates this clearly. I have then argued that by restricting the use of commercially developed "off the shelf" algorithms these problems can be avoided altogether.

4.3. Gaming the System

4.3.1 Introduction

If too much information about how an algorithmic system works is made available to potential data subjects they have the opportunity to alter their answers to input questions in order to make the output more favourable for them. In this section I will introduce the concept of gaming the system. I will then go on to discuss the problems that result from an algorithm being gameable. These include threats to accuracy and putting some data subjects at a disadvantage. I am going to argue that the problem of gaming the system only arises if the algorithm is using dynamic factors. I will conclude this section by considering what this means for the use of dynamic factors in risk assessment.

4.3.2 What is gaming the system?

"Gaming the system" is a term used to describe a situation in which someone manipulates the rules of a system in order to get a desired outcome (Gunkel 2018, 1). A good example of gaming the system comes from John Hopkins University in Baltimore, Maryland. The students of Peter Fröhlich, a computer science professor, were aware that his marking schedule was set up in such a way that the student who had performed the best would automatically get an A and then all the lower grades would be assigned accordingly. Instead of all going to the exam for the course the students collectively boycotted it. You would think that this would lead to every student in the class getting a zero in the exam. However, this was not the case as the students had realised that the marking schedule rules meant that they

could all end up with an A grade if no one attended the exam because the zero grade would technically be the highest grade in the class and would be awarded an A grade (Rampell 2013). Essentially, the students had seen a way in which they could manipulate the system and chose to do so because they could see the benefits they could accrue.

4.3.3 "Gaming" in the criminal justice system and its effect on accurate predictions

The likelihood of an offender being able to game the system is something that must be considered when it comes to the use of predictive algorithms. Too much transparency can be a gateway to a person effectively gaming the system if operators are not careful (Caplan 2018, 7; EPFL IRCG 2018, 22). Gaming is likely to happen in any case where an offender is dishonest and can see how to manipulate their answers for maximum benefit. As Cathy O'Neill, author of *Weapons of Math Destruction,* states that 'most of the prisoners filling out mandatory questionnaires aren't stupid. They at least have reason to suspect that information they provide will be used against them while in prison and perhaps lock them up for longer' (O'Neill 2016, 28).

The questions used in an algorithm such as COMPAS (available in Appendix C) have such an obvious valence that many ordinary people would see what changing their answers would do. An example of this valence in the COMPAS questionnaire is a question that asks an offender 'I am seen by others as cold and unfeeling', with possible answers of 'Strongly Disagree', 'Disagree', 'Not sure', 'Agree', 'Strongly Agree' (Northpointe 2016). Here an offender is likely to see that saying people see them as cold an unfeeling would make them look more antisocial and thus more likely to offend than if they did not, so they could alter their answer to make it look more favourable. This "gaming" problem would be exacerbated if the offender knew the weightings on each input for producing the output. This is because they would be able to make exact calculations about what their risk score would be if they

answered honestly versus what it would be if they changed their answers to try and decrease their risk score. They could then go on to answer the questions used as inputs with complete confidence that they could manipulate the risk score in their favour. The difference between knowing the weightings and not is that without knowledge of the weightings the offender would be guessing, whereas if the offender knew the weightings they could game the system with absolute certainty regarding how their answers would change the risk score.

Edwards and Veale claim that it seems 'unlikely that prisoners will change their characteristics just to attempt to game a recidivism algorithm that will not even be used until after they have been apprehended' (Edwards and Veale 2017, 63). However, this seems mistaken. It is mistaken to claim that people will not game the algorithm based on the fact that it will not be used until after the data subject is apprehended. A risk score is not developed until after the offender has been apprehended (Casey et al. 2014, 2). By this point it would be unlikely that the offender did not know that how they answer the questions will make a difference to their future. Furthermore, if the offender does know what the algorithm is used for, and how much impact it has on a decision, it is likely that the offender would have an interest in altering their answers. This would be more likely if it might stop them from receiving a longer sentence or not making bail.

As Andrew Guthrie Fergusson states, some of the system must remain unknown to the public in order that the system can continue to make accurate predictions (Guthrie Fergusson 2017, 9). The fact that an offender could change their answers in such a way that it gives them a different risk score than if they had answered the questions honestly is concerning. It is concerning because in cases where judges and parole boards are using the risk scores to inform their decisions the risk score is giving them information that does not reflect the risk of the offender re-offending. This may lead to judges and parole boards acting in a way that does not reflect the level of risk that the offender is actually at of re-offending.

4.3.4 Gaming the system and cognitive resources or honesty leading to a disadvantage

A possible problem that could follow from an algorithm that is gameable and that may motivate a call for more opacity is that those with fewer cognitive resources or those who are honest end up at a disadvantage[17]. If an offender can game the system by making calculations then there will be some people who have the cognitive resources to see how to do so and some who do not. Cognitive resources could include access to the internet, access to a lawyer, and an ability to problem solve well[18]. Also, there will be a number of people who, whilst able to game the system, are too honest and truthful to do so. So, another reason that gaming the system might encourage more opacity is that disadvantaging people based on their being honest, or having unequal access to cognitive resources is unfair.

Better cognitive resources are neither necessary nor sufficient for a person to game the system in the way I am suggesting. Rather, those with greater access to resources and a better ability to make the calculations to game the system are at an advantage here if weightings are made available. Those who can game the system might end up with a low risk score and that then plays into their getting treated in a more lenient manner than they would have been had they answered the questions honestly. On the other hand, those unable to see how to game the system, or those too honest to do so, will potentially receive harsher treatment than those who have acted dishonestly. This can cultivate the argument that operators and developers ought to make inner workings of algorithmic systems unavailable to data subjects and the public. If no one knows what effect the weightings of each input has

[17] "Cognitive resources" is not just a euphemism for intelligence. The term is in fact intended to capture more than just intelligence. Cognitive resources can also be external to the data subject and can include; access to a lawyer, a comprehensive education, and access to the internet.

[18] There are several other factors that would also add to a person being more likely to game. This point is simply to say that those with more knowledge tend to be at an advantage when it comes to gaming the system.

on producing an output then at least an attempt to game the algorithm is based just on guesswork.

There is an objection to this argument, which is that even without knowing the inner workings of an algorithm someone could effectively game it. The COMPAS algorithm serves as a good example here. COMPAS has 137 questions that are used as variables in the algorithm (Dressel and Farid 2018, 1) (See Appendix C for sample COMPAS questionnaire). Some of the questions are such that most people could see that answering them in a particular way would yield a particular result. In this case the person with more cognitive resources probably still has an advantage. However, systems such as COMPAS do not divulge the weightings of these variables or whether all the variables are used to make the calculation (Northpointe 2016). This approach at least decreases the disadvantage that a person with fewer cognitive resources might experience.

4.3.5 Gaming the system is only a problem with algorithms that use dynamic factors

In chapter two I discussed static and dynamic factors. As a refresher, static factors are factors about a person that they cannot manipulate. For example, their age, offending history, and days since last offence[19]. Dynamic factors, on the other hand, are factors that an offender could choose and may change over time. For example, whether the offender often feels bored, or whether they suffer from substance abuse. Gaming the system in the criminal justice case can only occur when the algorithm uses dynamic factors. Because dynamic factors are manipulable, a problem that results from using dynamic factors and having too much transparency is that an offender who understands the weightings placed upon the factors used in the algorithm might misreport their answers to certain questions in order to

[19]A static factor can change over time, for example a person's age. However, the offender has no control over what their age is.

receive for example, a lower risk score. This cannot happen if the algorithm uses only static factors such as the number of days since last offence, or offence history, because these are facts about the offender that the offender simply cannot change. Peter Johnston reiterates this point, stating that New Zealand's RoC*RoI algorithm is influenced only by static factors, which means they 'cannot be changed no matter how 'reformed' the person becomes' (Johnston 2018).

Whilst systems such as New Zealand's RoC*ROI cannot be gamed because only static factors are used some algorithmic systems use dynamic factors. This means that that, even without knowing how a question is weighted, an offender could potentially supply answers that do not reflect what is actually the case at that given time. I demonstrate how an offender might game an algorithm using the COMPAS algorithm as an example in §4.3.3.

4.3.6. Mitigating the negative effects for those who are honest, or have fewer cognitive resources

As I discussed in §4.3.4. those with fewer cognitive resources are put at a disadvantage if they are not able to see how to game the system. I then also pointed out in section §4.3.5., that gaming the system will not occur when using an algorithm that only uses static factors. As a result, gaming the system will not be a problem for an algorithm that relies solely on static factors to make calculations. However, it is something that still must be considered in cases where dynamic risk factors are also used to calculate a risk score. In order to avoid too much disadvantage operators should make the weightings on the algorithm unavailable to data subjects (and the public). This will not entirely get rid of the disadvantage but is the best that can be done for now.

4.3.7 Transparency recommendation

The "gaming the system" justification only holds in the dynamic factors case, as static factors cannot be manipulated by the offender like dynamic factors can. In the dynamic factors case there is a tension. It is important to maintain opacity so that the algorithmic system can remain as accurate as possible, yet, in order to reduce public doubt, give offenders a chance to give informed consent, and allow them to utilise their right to appeal a decision, it looks as though many of the details should be made available to the public. It looks then like the problem of gaming the system cannot be completely removed by making the entire algorithmic system opaque. However, making the weightings unavailable may go some way to avoiding the problem

Another suggestion to mitigate the gaming problem would be to remove dynamic factors from being used in predictive risk models. Doing so would remove the problem of gaming the system entirely. However, doing so would take away a major benefit that using dynamic factors can bring to a risk prediction model. Using dynamic factors allows the operators to see changes over time and offer rehabilitative programmes that are tailored to the offender (Ward and Fortune 2016, 80). Whilst it is possible to construct relatively accurate algorithms without dynamic factors it would be more advantageous if controls concerning the use of them could be developed, in order that the benefits of using dynamic factors could still be reaped.

Perhaps dynamic factors should only be used in cases where the risk score produced by the algorithm helps a decision-maker assign rehabilitation services to an offender. This appears to be the context in which dynamic factor use is most beneficial because it shows changes in the offender over time (Ward and Fortune 2016, 80). This would also mean that people would not game the system in cases where parole and sentences were involved, which is a far higher stakes context for gaming to occur. However, this seems like a radical

approach. Instead weightings of inputs used in producing outputs should not be available in the dynamic factor case, and regulations should require that, where dynamic factors are used any *ex post* explanation must provide the most prominent factors (rather than the weightings) that led to an individual's risk score.

4.3.8 Concluding remarks

In this section I have shown that in cases where dynamic factors are used there is a risk of an offender "gaming the system". This can lead the algorithm to produce a risk score that does not reflect the risk of the offender re-offending. Furthermore, it can disadvantage those who cannot or will not game the system. I have argued that in cases where dynamic factors are used that it is important to not reveal the weighting of inputs in producing outputs. Instead of revealing weightings the offender can be given the main factors that lead to their risk score.

4.4. Too much transparency can reduce trust in system

4.4.1 Introduction

The interaction between public trust and transparency is interesting and impacts the way in which transparency is thought about in the context of algorithms. Some level of transparency is thought to increase trust by 'building credibility' but it appears to be a balancing act, as too much or too little transparency can in fact lead to a lack of trust (Heald 2006a, 62).

4.4.2 How can transparency threaten trust?

People do not trust black box algorithms, but they also do not trust an algorithm that they perceive to be inaccurate, or that uses facts to make predictions that they do not agree with.

A reason to call for more opacity is that it could decrease the doubt people have as a result of perceived (but not actual) inaccuracy or bad fact choice. By not making all details about the inner workings of an algorithm available it is possible to decrease the likelihood of the public trying to axe an algorithm that works far better than a human completing the same task. Hannah Fry states that humans can go from over-trusting an algorithmic system to wanting rid of it very quickly often without allowing any proper assessment of how the algorithm is functioning (Fry 2018, 64). So, perhaps, by keeping algorithm factors or their weightings hidden an operator may be able to continue using a perfectly good algorithm without public doubt leading to a lack of support (Fry 2018, 64).

4.4.3 Kahneman Analogy

Daniel Kahneman states in his book *Thinking Fast and Slow* that it is often an illusion that leads us to trust processes (Kahneman 2011, 217). He uses the example of stock markets to demonstrate this point. He claims that hardly anyone can consistently produce good results year after year in the stock market (Kahneman 2011, 217). Those working in the stock markets essentially present the public with the illusion that buyer and seller have skill when it comes to selling and purchasing stock and this keeps them happy and trusting of the system (Kahneman 2011, 217). The illusion of well-founded prediction gives the public the impression that the information is 'privileged, or at least extremely insightful' (Kahneman 2011, 218). Which is to say that sometimes it is best to keep details hidden. In this case it is because the public will no longer trust the system if developers and operators are completely open about how it works.

Something similar could happen in the algorithm case. Sometimes an operator might think that keeping certain elements of an algorithm hidden would help to maintain trust. For example, imagine it is found that an offender's income is a highly predictive factor and using

it makes recidivism predictions far more accurate. As a result the developers of the algorithm put a heavy weighting on this variable. Imagine then, that the public disagree that this factor should be used in a predictive risk assessment. They may, as a result, begin to withdraw their support for the algorithm to be used, as they no longer trust it to be accurate. It might then be useful for the operators and developers to withhold information about income being highly predictive factor and that it has a heavy weighting in producing the risk score. By not sharing this information trust might be maintained and a more accurate prediction achieved. However, this is not a good reason for withholding factors from data subjects. This is because it may be morally wrong to use certain factors and by keeping them opaque the operators may end up using morally questionable factors without the public knowing (Kehl et al. 2017, 28).

Another way in which doubt may arise is the public not understanding how the criminal justice algorithm works or is operated. This may lead them to draw incorrect conclusions about the algorithms if they are given too much complex information about the algorithm. Instead, it might be more helpful to provide a simpler and less detailed explanation about how the algorithm works and how it is operated. Offering the public direct access to the inner workings of a criminal justice algorithm would potentially be too confusing and would probably provide them with information that is neither useful nor utilisable.

4.4.4. How to maintain trust

It is difficult for a regulator to determine what sort of doubts the public might have prior to the introduction of an algorithm into the criminal justice system. However, regulators can probably make a good guess. Given the evidence above they can probably guess that keeping an algorithm completely opaque would lead to data subjects and the public doubting the good that the algorithm is doing. They could also guess that being too transparent might lead to

citizens forming the wrong impressions due to not understanding the information made available. However, there is opportunity to get these guesses wrong and therefore the public must be able lodge a complaint and ask for review of the degree of transparency.

Opportunities should also be put in place to review and change the level of transparency if great doubt in the system grows. For example, by having a digital ombudsman with whom to lodge complaints. The Government also 'must understand the need to build public trust and confidence in how to use artificial intelligence…' (Select Committee on Artificial Intelligence 2018, 25). If a government looks to be open to review as well as open with information then citizens will feel more trusting as they will feel their government is not trying to hide things from them. Furthermore, effort must be put into developing explanations that a lay-man could understand. There is little point in providing either a data subject or the public with information that they cannot utilise.

4.4.5 Concluding remarks

In this section I covered the problem that sometimes revealing too much information to the public can lead to a decrease in trust rather than an increase. I suggested that in some cases it might be worth keeping things opaque in order that trust is maintained. The first case I covered was when factors were being used that the public might disagree with. I claimed that this was not an appropriate reason for opacity. The second case I covered was in cases where technical information was given to the public and they leapt to incorrect conclusions. Here I suggested that simplified explanations might fill this gap. Finally, I suggested that in order to maintain trust regulators must do their best to set rules that bring about trust and I further suggested that opportunities for review were put in place, and finally I suggested that simplified explanations might be given where possible.

4.5. Transparency limits due to complexity

4.5.1 Introduction

In chapter two I touched on the idea that in some cases there are technical limits to developers and operators being able to explain how an algorithm works. In this section I will explain that this can inhibit a developer's and an operator's ability to be transparent. I will then assess what options we have to get around this problem.

4.5.2 Conflict between requirements for transparency and technical limits

In some cases the way that an algorithm works is so complex, or there is so much data, that important aspects of how an algorithm works cannot be shared. For example, HART has 4.2 million decision points, and developers of artificial neural networks often don't understand how their own creations work (Oswald et al. 2018, 12; New and Castro 2018, 5). Whilst it is a logical possibility that every single factor could be reported, it would be extremely time consuming, difficult, and probably not very useful to the data subject. It can be a problem if a data subject has rights that require a certain level of transparency and the algorithm is too complex or too big to be able to give the information needed. In chapter three and four I have discussed what is required in terms of *ex ante* and *ex post* transparency to develop and operate a predictive risk algorithm ethically in the criminal justice system. The problem is that there will still sometimes be limits to transparency. This final section will briefly look at what options there are when this occurs.

4.5.3 Simplified explanations

If an algorithm is too complex to understand to begin with, the developer and the operator should attempt to give simplified explanations to data subjects and citizens. This simplified explanation should be sufficient to explain details to a data subject such that they are able to

give informed consent and exercise their right of appeal. Explanations should also give enough information to the public such that they do not doubt the use of the system and are able to maintain control through accountability. In cases where it is not possible to achieve this there are two other options to consider (i): a model of a model approach and (ii) not operating the algorithm at all.

4.5.4 Model of a Model

One approach that Edwards and Veale suggest is that a model of a model can be used to explain how an algorithm is determining its outputs and to reassure developers, operators, and data subjects that the algorithm is working as anticipated (Edwards and Veale 2017, 54). Ribiero et al. describe this model of a model as an algorithm that can 'explain the predictions of *any* classifier or regressor in a faithful way, by approximating it locally with an interpretable model' (Ribiero et al 2016, 1). This approach would be useful in cases where even the developer does not understand what the algorithm is doing. The idea is that an extremely complex model could have a second model applied to it that could present 'textual or visual artifacts [by-products produced during development] that provide qualitative understanding of the relationship between the components of the algorithm and the prediction (Ribiero et al. 2016, 1). Such an approach would extract an explanation without taking apart the model (Edwards and Veale 2017, 54). However, there are a few problems with using the model of a model approach. Firstly, what is to say that the second model (the model of a model) can be any more trusted than the first model? Secondly, and more importantly there is very little literature to support the effectiveness of this approach and computer scientists such as Cynthia Rudin argue that these 'explanations are often not reliable, and can be misleading' (Rudin 2018, 1). So, at the time of writing, a fair amount

more research (and evidence of success stories in less high stakes situations) needs to occur before the model of a model approach is tested, and then used in a criminal justice system.

4.5.5 In cases where a simplified explanation cannot be given

It looks at this point as though a model of a model approach is not an appropriate approach. Given that this is the case there is only one option left if a simplified explanation cannot be provided. To make sure that technical limits do not infringe on required transparency, if a simplified explanation that contains the required information cannot be given, then the algorithm should not be used.

4.5.6 Concluding remarks

In this section I have explained that sometimes the algorithm is too complex to explain in full. I explained that the model of a model approach to providing information about the algorithm is too new and untested to help us overcome this problem. I have also claimed that if an explanation could be simplified and generalised in such a way that the data subject could still, for example, utilise their right to appeal then an algorithm could still be operated. However, if a simplified explanation cannot be given the algorithm cannot be used.

4.6. Conclusion

This chapter has looked at four main reasons that transparency might be limited or have to be limited. These included trade secrets, gaming the system, public trust, and technical limits. Where possible, alternative approaches have been given in order to remove transparency limits. At this point both requirements for transparency and limitations of transparency have been identified. These will be used in the following chapter to discuss the way in which transparency of algorithms can be regulated.

5.0. Chapter Five: A Transparency Framework for operating criminal justice algorithms in New Zealand

5.1. Introduction

Due to considerations such as the importance of consent, a right to appeal, and the need to maintain accountability certain elements of an algorithm need to be made available to data subjects or the public. We must construct a framework that specifies what transparency requirements a criminal justice algorithm must meet for it to operate in the New Zealand criminal justice system. In this chapter I will stipulate a number of requirements that must be observed.

I will look at three ways in which people have attempted or proposed to control the transparency of algorithms around the world. I am going to begin this section by assessing whether an overarching body should regulate algorithms or whether each algorithm needs to be independently assessed. From here I will look at the European Union's General Data Protection Regulation (GDPR) and argue that it does not provide an appropriate way of securing the desired level of transparency. I will then look at the ALGO-CARE framework and claim that they do not cover what is needed for regulation nor are they specific enough. Finally, I will look at New Zealand's 'Right of access by person to reasons for decisions affecting that person'(*Official Information Act 1982* , sec. 23), and argue that it does secure some transparency but not enough.

Having shown that these approaches either fail to capture the transparency requirements at all, or that they do not capture *all* requirements I will move on to proposing my own framework. I acknowledge that transparency is not the sole way to regulate and that by just focusing on transparency I will be missing important elements that should also be regulated. However, transparency is very diverse and is an instrumental way of achieving

other objectives (Heald 2006a, 59). Requiring transparency in certain ways can lead to several other important requirements being met. I think, therefore, that regulating levels of transparency is an important place to start. Further requirements can be added as more research is done about what transparency regulations alone do not achieve.

5.2. What sort of regulatory body do we want?

Andrew Tutt, concerned by the harms posed by using algorithms and the fact that that they pose a 'unique regulatory challenge', suggests that governments have something like an FDA for algorithms (Tutt 2017, 91). The idea is that there would be an over-arching regulatory body which would assess all algorithms before they were used. This is a concept similar to the FDA which tests drugs before they can be given to the public. Tutt claims that there are a few benefits to this approach (Tutt 2017, 105).

Firstly, Tutt claims that an overarching agency can act as a standard setting body. Standards can be set regarding how much scrutiny the algorithm should come under and what performance, design, and liability standards the algorithm needs to meet (Tutt 2017, 105–9). Secondly, Tutt claims that a regulatory body could act as a soft touch regulator, which is to say that they could 'impose regulations that are low enough cost that they preserve freedom of choice and do not substantively limit the kinds of algorithms that can be developed or when or how they can be released' (Tutt 2017, 109). These regulations could include transparency regulations about openness and disclosure (Tutt 2017, 110). Thirdly, Tutt also proposes that they could act as a 'Hard-Edged' regulator which is to say that they could impose heavy restrictions on the use of algorithms. For example, they could require pre-market approval before algorithms are used (Tutt 2017, 111). A big advantage of having an external regulatory body is that it is a method of enforcing the assessment of *all* algorithms.

Despite the surface appeal of this proposal, having a general regulatory body that specify rules for all algorithms is problematic. The most obvious reason why this is not the best approach is that algorithms tend to be used for very specific purposes. In many cases such as the criminal justice system where the stakes are high for data subjects, it requires a close understanding of the purpose to properly analyse whether the algorithm is appropriate. Algorithms, particularly those which have the potential to have a big impact on a data subject need to be considered with context in mind. The worry is that an overarching regulatory body would not have knowledge that was specialised enough to complete this task.

An overarching set of general rules and an overarching method for assessing an algorithm is not appropriate. Different algorithms require different sets of considerations. The requirements needed for the operation of a Spotify algorithm (that is very unlikely to cause a person harm) are radically different from the requirements needed for the operation of a criminal justice algorithm (where the harm can be greater). Furthermore, to assess the algorithm in a way that is detached from the system which it will operate is a sure-fire way to make recommendations or requirements that do not work in practice. This is because there are many different applications of algorithms, and it will not be helpful, for example, to regulate the use of a criminal justice system in the same way as the algorithm that New Zealand's Ministry of Education uses for planning bus routes, and calculate a child's eligibility for transport assistance (NZ Stats 2018, 13)

The best approach is to have an overarching body that requires that each algorithm be assessed and each area in which an algorithm operates have a stipulated set of standards that have to be observed. However, it is important to require that the standard setting must be left to an independent expert in that algorithm and the context that it is used in. The standards that are set need to be customised to the context within which an algorithm will operate. Whilst an overarching body will be helpful in ensuring that standards are, in fact,

set, these standards should be set by an independent assessor with a thorough understanding of the context within which an algorithm will operate. Once the independent assessor has developed these standards they could then be passed to an overarching regulatory body that can approve or reject the standards built by the independent assessor and formalise them as regulations. Those in the business of operating or developing algorithms once this is done should have to provide proof (and ongoing proof at that) to the overarching body that they are following the regulations or set of rules set out for their specific department.

5.3. Assessing attempts at transparency regulations

As algorithms have become more commonplace it has become more obvious that governments ought to regulate the use of them. Many governments, academics, and independent researchers are thinking about how best to do this (Select Committee on Artificial Intelligence. 2018; Goodman and Flaxman 2017). In this section, I will look at three approaches to regulating algorithms, and whether they secure the transparency that I have determined is important in chapters three and four.

5.3.1 GDPR

One approach to the regulation of algorithms is the GDPR which was enacted in May of 2018 and which replaced the Data Protection Directive 1995 (DPD) (Directive 95/46/EC 1995). The purpose of the GDPR is to set a standardised set of laws across the European Union to help protect citizens when data belonging to them, or about them is used (European Commission 2018). The GDPR will regulate how data is used and make it easier for citizens of the European Union to understand how their data is used (European Commission 2018). Lee Bygrave states that in the past the European Union's data protection law has been seen as a model to follow as many think it provides 'a qualified right for a person not to be subject

to fully automated decisions based on profiling and supplements this with a right to knowledge of the logic involved in such decisions' (Andrews et al. 2017, 31). However, the GDPR has several weaknesses so this may no longer be the case.

Lawyers and academics alike have looked for a right to an explanation in the GDPR which could secure a sort of ex ante transparency (Goodman and Flaxman 2017; Edwards and Veale 2017; Andrews et al. 2017, 31-34; Wachter et al 2016.) However, there is disagreement about whether the GDPR achieves this or not (Goodman and Flaxman 2017; Edwards and Veale 2017; Andrews et al. 2017, 31-34, Wachter et al. 2016). Article 22 is the most obvious place to start when it comes to possibly finding a right to explanation for algorithms in the GDPR. Article 22 states that a 'data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her' (*General Data Protection Regulation* [2018], Article 22). On its own this section only gives a person a right to not have an automated decision made that affects them. However, Recital 71, set out to explain Article 22 states that:

> [S]uch processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision
>
> (*General Data Protection Regulation* [2018], Recital 77)

This could possibly guarantee some sort of "right to explanation", but there are two major problems here. Firstly, a recital is not legally binding. A recital is only there to

disambiguate what the article it is related to says. Unfortunately, as Lee Bygrave states, 'clumsy syntax … muddies [Article 22's] interpretation' (Andrews et al. 2017, 32–33). So it is not obvious that there is a way to give a right to explanation in Article 22.

Secondly, this right only covers decisions that are fully automated. The consequence of this is that this article will not cover criminal justice algorithms because the outputs of the algorithm are simply used to help aid decisions not make them. Risk Prediction Algorithms used in justice systems such as COMPAS, RoC*RoI, and HART would not be covered by the GDPR as their outputs do not meet the 'fully-automated' standard.

At this point it looks as though Article 22 and Recital 71 are a lost cause when it comes to securing a right to explanation at least in the criminal justice algorithm case. However, even if they did achieve a right to explanation for algorithms which have human input this does not look like it will provide the transparency that is determined to be important in chapters three and four. This is because the GDPR (if it did secure a right to explanation) would only guarantee some form of *ex post* transparency and as I determined in chapter four, there are strong arguments for *ex ante* transparency.

Edwards and Veale (2017, 53) suggest that Article 15 might be a better way of securing a right to an explanation than article 22. Article 15 stipulates that:

> The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:
>
> 1. the purposes of the processing;
> 2. the categories of personal data concerned;

3. the recipients or categories of recipient to whom the personal data have been or will be disclosed, in particular recipients in third countries or international organisations;
4. where possible, the envisaged period for which the personal data will be stored, or, if not possible, the criteria used to determine that period;
5. the existence of the right to request from the controller rectification or erasure of personal data or restriction of processing of personal data concerning the data subject or to object to such processing;
6. the right to lodge a complaint with a supervisory authority;
7. where the personal data are not collected from the data subject, any available information as to their source;
8. the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

(*General Data Protection Regulation* 2018, Article 15).

This looks like a more straightforward approach to achieving a right to explanation, especially as it does not discount algorithms where there is human input. However, Edwards and Veale claim that this approach still has problems.

Article 15 gives the data subject rights after their personal information has been put into the system, which has the upside of the data subject receiving the 'logic, rationale, reasons, and individual circumstances of a specific automated decision' (Select Committee

on Artificial Intelligence 2018; Goodman and Flaxman 2017; *General Data Protection Regulation* 2018, Article 15). This looks very straightforward and a good way of securing an explanation. However, it does not guarantee that the explanation will be useful or in an easily understandable form for the data subject. If it is not in an understandable form access to the information seems useless.

Furthermore, Recital 63 allows for the protection of trade secrets and other forms of intellectual property when it comes to observing Article 15. The implication of this is that an algorithm developer could stop an offender from receiving an explanation if it might give away their trade secrets. As I have determined in chapter four trade secrets should not come in to the equation when it comes to government's operating algorithms that impact their citizens.

Edwards and Veale argue that it is difficult to see how the GDPR secures transparency. Furthermore, they say even if it were to secure transparency it might well be 'meaningless transparency' (Edwards and Veale 2017, 23). Wachter et al. hold an even stronger view, which is that the GDPR gives no right for transparency in non-automated decision making (Wachter, Mittelstadt, and Floridi 2017, 78). Even if Article 15 did secure some sort of explanation though the GDPR would only provide one aspect of transparency that should be required for the operation of an algorithm. The GDPR fails to be specific enough to regulate predictive risk algorithms that are used in the criminal justice system (this is why requirements should be set with context in mind). As I determined in chapters three and four the operation of criminal justice algorithms also require an *ex ante* transparency which the GDPR does not achieve[20]. Therefore, the GDPR is not an appropriate tool for

[20] Wachter et al. state the GDPR does guarantee an ex ante transparency but only for fully automated decisions, which criminal justice algorithms are not. Even then, the requirements are very narrow. For more information Wachter et al. discuss it in *Why a Right to Explanation of Automated Decision-Making Does Not exist in the General Data Protection Regulation (Wachter, Mittelstadt, and Floridi 2017).*

regulating a predictive risk algorithm that is to be operated in the criminal justice system. Furthermore, how algorithms are used in criminal justice is very place and context specific, so we also need the regulations to be developed with this in mind.

5.3.2 ALGO-CARE

ALGO-CARE is a framework developed by Oswald et al. that lists several considerations that reflect the experience of those that developed the HART algorithm in County Durham in England. The aim of the ALGO-CARE framework, that they have started to develop is to 'translate key public law and human rights principles into practical considerations and guidance that can be addressed by public sector bodies' (Oswald et al. 2018, 244).

The ALGO-CARE framework is meant to be a 'guide' to regulation rather than propose regulations. ALGO-CARE is still in the development phase so it is unfair to criticise it as though it is being used for regulation. Oswald et al. recognise that people might interpret it as an attempt at developing regulations and so state that they 'appreciate that this framework does not provide any firm answers' (Oswald et al. 2018, 249). In this section I will assess whether it at least acts as a good guide for the regulation of transparency in operating criminal justice algorithms.

The ALGO-CARE framework has many points that Oswald et al. have identified as being important when it comes to considering how to regulate a risk prediction algorithm like HART (the ALGO-CARE framework can be found in Appendix D). The points that relate to transparency include:

1. 'Who owns the algorithm and the data analysed? Does the force need rights to access, use and amend the source code and data analysed?'

2. ‘Are there any contractual or other restrictions which might limit accountability or evaluation?’
3. ‘Where an algorithmic tool informs criminal justice disposals, how are individuals notified of its use (as appropriate in the context of the tool’s operation and purpose)’[21]
4. ‘Is the use of the algorithm transparent (taking account of the context of its use), accountable and placed under review alongside other IT developments in policing?’
5. ‘Is appropriate information available about the decision-making rule(s) and the impact that each factor has on the final score or outcome?’
6. ‘Is the force able to access and deploy a data science expert and justify the algorithmic tool (in a similar way to an expert forensic pathologist).’

(Extracted from Oswald et al. 2018, 245–48)

The ALGO-CARE voices many considerations that that have been determined as necessary in chapters three and four. These include whether there is enough information made available about the decision-making and factors used, whether those involved have access to source code, whether there are restrictions that will hinder access, and whether someone or somebody is accountable for the use of the algorithm. These are all important considerations that any person building and operating an operating should be taking into

[21] By “criminal justice disposals” Oswald et al. mean “a way of dealing with an offence not requiring prosecution in court)” (Oswald et al. 2018, 225).

account. However, this list should include more considerations of the information that should be available to data subjects. For example, it does not guide us to consider what information needs to be made available for data subjects to utilise their rights such as a right to ask for a review, or a right to appeal.

The ALGO-CARE framework also does not give enough consideration to the need for *ex ante* transparency. The only *ex ante* transparency that it recommends is that subjects know that an algorithm is being used. As determined in chapter four those operating criminal justice algorithms need to allow subjects to give informed consent for an algorithm to be used and this requires *ex ante* transparency. Furthermore, *ex ante* transparency is important for achieving public trust (though quantifying what level of transparency is difficult).

To be of use ALGO-CARE needs more detail than it currently has. This is a somewhat unfair criticism as Oswald et al state that it is still in a development phase (Oswald et al. 2018, 249). However, it is important to recognise that in order to regulate there need to be specific and thorough thresholds that must be met. ALGO-CARE is only a guide to regulation so as I have stated previously it is unfair to treat it as anything else. However, as it stands it does not give us much direction. Those who use it is as a guide are essentially given a list of items they then should 'consider'. But not actually what they should have to take into account when forming regulation.

In order for the ALGO-CARE framework to properly guide regulation it would be useful for it to have slightly more concrete considerations including some *ex-ante* ones. It would also be more beneficial if there were some more 'data subject – centric' considerations. At the end of the day whilst these considerations are helpful they are simply too vague. Those forming regulations for algorithms might take the considerations that Oswald et al. have shared into account when starting to think about what they might need to regulate but that is the extent of its use.

5.3.3 Official Information Act: 'Right of Access by person to access reasons for decisions affecting that person'

Looking at regulating the use of criminal justice algorithms in New Zealand it is worthwhile turning to legislation that already exists to see if it might be used to regulate transparency in some way as it may be less time-consuming or costly. The most promising right in New Zealand that might guarantee some sort of transparency for a data subject is in the Official Information Act 1982 (OIA). This right is the 'Right of access by person to reasons for decisions affecting that person' (*Official Information Act* 1982, sec. 23).

> The 'Right of access by person to reasons for decisions affecting that person' (*Official Information Act* 1982 (23)) states that a person can receive a written statement of:
>
> 1. the findings on material issues of fact; and
> 2. subject to subsection (2A), a reference to the information on which the findings were based; and
> 3. the reasons for the decision or recommendation.
>
> (*Official Information Act* 1982, sec. 23)

However, there are certain limitations that could come into conflict with exercising this right, one of which is particularly relevant if an algorithm is a commercial algorithm. Reasons can be withheld to:

> protect information where making available of the information-
>
> 1. would disclose a trade secret

2. would be likely unreasonably to prejudice the commercial position of the person who supplied or who is the subject of the information

(*Official Information Act* 1982, sec. 9(2))

This shows the need for additional regulations about trade secrets and, also, that the Official Information Act 1982 (23) will not do all the work that it needs to.

This right to reasons provides some sort of ex post transparency. It can ensure that an individual gets an individualised set of reasons about how a risk score that is produced by a criminal justice algorithm was used to make a decision. A question arises here about whether it would be able to secure reasons for why the inputs led to the output. The current law allows 'for reasons for the decision or recommendation' (*Official Information Act 1982*, sec. 23(3)). However, this might not include reasons for how the algorithm developed the risk score that was then used to help make a decision about the offender.

As I determined in chapter 4 *ex ante* transparency is required in order that an offender can give informed consent for a risk score to be developed in cases where the offender is asked to provide information, as well as to build and maintain trust. The right for reasons does not secure this as it only acts as a form of *ex post* transparency (and it is limited). It also only provides information to the offender about a decision made that affects them which means it does not guarantee any transparency between developer and operator nor does it guarantee any transparency with the public, both of which I determined were important in chapter four. Furthermore, it is important to have something in place to make sure that trade secrets, which restrict transparency regarding the weightings of inputs in producing outputs and other similar elements, do not infringe on an individual's rights concerning the use of a risk of recidivism algorithm.

5.4. A transparency framework for criminal justice algorithms in New Zealand

What I have shown by looking at the approaches above is that they are not achieving enough transparency or the right sort of transparency. In this section I will specify several requirements relating to transparency that will help to achieve fair operation of an algorithm. These including: removing roadblocks to transparency, transparency limits based on the factors used, transparency between developer and operator, transparency required to observe rights and maintain public trust, and dealing with technical limits to transparency. Of course it is important to regulate more than just transparency when it comes to using algorithms. However, as many scholars also suggest, transparency is the best place to start (Tutt 2017, 110; Pasquale 2015, 140–88; Mayer-Schönberger and Cukier 2013, 176–84).

5.5. Transparency Framework

Removing Road Blocks to Transparency

Trade Secrets

Algorithms should be developed in house, or through a contract which requires operator access to source code. This approach stops developers from restricting access to source code, and other inner workings of an algorithm.

Limits on Transparency based on factors used

Dynamic Factors

Where an algorithm uses dynamic factors adjustments should be made to prevent gaming of the system. Weightings of each input in producing the output should not be disclosed to the data subject. *Ex ante* explanations can be given but should not include source code, or information about how the algorithmic system is weighted. An *ex post* explanation can also

be given, but only general reasons for the output should be given. For example: the main factors that lead to the offender receiving that risk score.

Static Factors

In cases where an algorithm only uses static factors weightings do not need to be withheld if it is determined that making the weightings available will be useful. However, when these details are made available a report is required that puts the information in an understandable form for the data subject, or someone who is representing them.

Transparency between developer and operator

Developers and operators should be completely transparent with each other. Developers should make available several things to the operator. These include: the source code, weightings, factors, and justifications for the choice of model. Operators must be clear with developers about the use of their algorithm. Developers and operators should work together and educate each other on both how the algorithm works and how it is operated.

Transparency needed to observe rights and maintain trust

Right to Appeal

Factors used to develop a risk score should be made available to the offender (though no weightings can be shared if the algorithm uses dynamic factors). The decision maker also must explain, if asked, how the algorithm was used in the decision-making process.

Post appeal or review request the appellant must receive a report explaining why the appeal was rejected or accepted. In cases where either the appeal is made on the use of the algorithm, or the appeal discovers that the risk score produced by an algorithm was unfair in

some way, then the data subject should receive an explanation of how it impacted the decision made.

Informed Consent

In cases where an offender is asked for information additional to information that the Department of Corrections or Parole Board hold on them they must give informed consent for that information to be used. They can choose not to provide the requested information. In order to make sure that they are appropriately informed there should be some *ex ante* transparency. In particular, they should know that the information that they provide will be used to make a decision about them, as well as the risks and benefits of providing this information. The argument in favour of this is that we allow people to refuse to give information in other cases, so we should do the same in this case.

Identifying those accountable

A person or group should be identified as accountable prior to the algorithm being used. Whilst this information need not be made available to the public, agencies developing and using algorithms should keep record of who is accountable for what, so in cases where for example, an appeal is made, there is a person identified as accountable for whatever part of the algorithm process is appealed, whether it be the construction, the operation, or use in decision making.

Public information in order to motivate accountable actors

When accountable, a person must be able to give an explanation or justifications for actions taken, or judgments made. To maintain accountability, the public should:

a) Know what the algorithm is being used for,

b) Have the main inputs that lead to the risk score made available to them,

c) Be presented with (a) & (b) in a way that they (or those representing them) understand.

Maintaining public trust

In discussions with experts in both the Department of Corrections and a potential regulator, a number of accuracy requirements need to be decided upon. The public should then be able to access information about whether the operator and developer are meeting these requirements.

The public must be able to lodge a complaint with an independent body (e.g. a digital ombudsman) about the way in which the algorithm works or is used. Transparency requirements should be included in such reviews. Alterations to the details shared with the public should be made if necessary.

Dealing with technical limits on transparency

If an algorithm is so complex that a simplified explanation cannot be given that satisfies transparency requirements for rights such as a 'right of appeal' to be utilized cannot be given, then an algorithm should not be operated.

5.6. Transparency framework in practice

These requirements are relatively stringent and one might object to this framework for that reason. As Chris Bousquet and Stephen Goldsmith state, regulators of algorithms need to be careful to not make the requirements so hard to reach that it leads to a return to no algorithms (Bousquet and Goldsmith 2018). The recommendations for regulation made above are

stringent, but the research in this thesis has suggested that it must be so in order that when a criminal justice system is operating a criminal justice algorithm it is does so ethically.

5.7. Why this framework only applies in New Zealand

This framework is specifically developed with the New Zealand criminal justice system in mind. Whilst many of the requirements would likely work for other countries it is dangerous to say that these requirements would apply in all other cases. Particularly as it relies on New Zealand legislation throughout. It is important that these requirements are customised to the algorithms which are used, the way in which the algorithms are being used, and the governments they are operating within. This is a complex and time consuming approach, but doing less would be unethical.

5.8. Conclusion

The GDPR, ALGO-CARE, and a 'Right of access by person to reasons for decisions affecting that person', do not ensure the transparency required for criminal justice algorithms. This is largely due to the lack of specificity and lack of *ex ante* transparency, that has been determined to be required for algorithm operation in New Zealand. Instead I have proposed a specific framework for regulating the transparency of predictive risk algorithms in the New Zealand criminal justice system. This framework achieves several things. Firstly, it removes road blocks to transparency. Secondly, it provides limits on transparency based on whether the factors used are only static, or if they are also dynamic. Thirdly it gives access to information between the operator and developer. Fourthly, it requires the transparency needed to observe rights and maintain trust. Finally, it directs operators and developers on what to do when the algorithm is too complex to explain.

Conclusion

At the beginning of this thesis I state that predictive risk algorithms that are operated in criminal justice systems around the globe need to be transparent in certain ways in order to operate well. In closing I will survey each chapter and its central purpose. Following this, I will address the part that transparency plays in the ethical regulation of criminal justice algorithms as well as point to future work that needs to be done. In the introduction, I make the claim that the use of risk of recidivism algorithms has raised questions about ethical practice and how to regulate algorithm use. My central questions were: are there ethical arguments for algorithms to be transparency? Are there legal arguments? If so, in what ways should algorithms and their operation be transparent?

In chapter one I provide an introduction to predictive risk algorithms used in the criminal justice system and explain why transparency in the operation of these algorithms is important. Firstly, I explain that in many jurisdictions around the world criminal justice algorithms are used to make predictions about whether an offender is likely to re-offend. These criminal justice algorithms produce a risk score, which is then used alongside other information to help decide about a prisoner's journey through the justice system. I explain that there are a number of reasons that these algorithms are used and that they add value that a human cannot. At this point I introduce three risk prediction algorithms that are used in various jurisdictions around the world. I contend that there is a need to ensure that algorithms are operated well. I claim that transparency is a way in which we can achieve this. I subsequently provide a number of popular justifications for transparency of criminal justice algorithms. These justifications include: public trust, accountability, appealing a decision or asking for a review, and informed consent.

In chapter two I further set the scene by explaining how predictive risk algorithms work, in particular, the type of assessment they offer and the models that they use. I start by introducing generations of risk assessment, focusing on second to fourth generation models as they are computer driven. From here I introduce static and dynamic factors and discuss the strengths and weaknesses of using them. I then explain what an algorithm is, and how algorithms learn. I also gave a summary of three common models used to predict risk. These are: random forest models, regression analysis models, and artificial neural networks[22]. With the scene setting done, the remainder of the thesis is an enquiry as to the amount of transparency required to operate criminal justice algorithms well and how we can regulate it.

In chapter three I revisit the justifications for transparency that were given in chapter one. I look at each of these in detail and argue that they are necessary elements of operating a criminal justice algorithm well. I then argue that transparency in some form will act as an instrumental way of achieving these important elements which include public trust, accountability, rights of appeal and review, and informed consent. At the end of each section I suggest the degree of transparency that each of these elements require. I finish this chapter by arguing that, in addition to the information that the public and data subjects must receive, there must also be complete transparency between developers and operators to increase the quality of the algorithm and to help ensure that good explanations can be given to data subjects.

In chapter four I claim that there are considerations that may lead to limiting transparency in some way. I look at these considerations in detail and assess whether they

[22] As I stated in §2.6.4. as far as I know there are no predictive risk models that utilise artificial neural network in order to develop risk of recidivism scores. However, they are worth mentioning as artificial neural networks are more accurate and are becoming more commonly used in other forms of risk assessment, such as likelihood of patient admission to hospital (Australian Government Department of Health 2019).

do, in, give good reason to limit transparency. Firstly, I consider the justification of trade secrets. This justification is given by some developers who will not make their algorithms transparent in case they lose their competitive edge. I argue that trade secrets should not be a reason for withholding information from a data subject as it conflicts with rights to appeal. I then claim that we can avoid the problem of trade secrets by requiring that an algorithm be developed in-house or by contract that requires all elements of the algorithm's construction are made available to the operator. Secondly, I argue that making criminal justice algorithms too transparent can lead to offenders "gaming the system", which can lead to decreased accuracy. I also argue that some offenders are disadvantaged due to being unable to see how to game the system, or are too honest to do so. I explain that "gaming the system" is only a problem in cases where dynamic factors are used to make a prediction. I suggest that in order to avoid this problem weightings of an algorithm should never be shared, and ex post explanations should be limited to the main factors that lead to the risk score. Thirdly, I argue that too much transparency can in fact lead to less trust. This also gives us good reason to not make algorithms entirely transparent to data subjects and the public. I suggest that in order to make sure that in cases where the public or data subjects question the transparency of an algorithm they can lodge a complaint with a representative such as a digital ombudsman who will review the transparency requirements. Finally, I explain that in some cases algorithms are very complex and not easy to understand. I recommend that in cases like this if a simplified explanation cannot be given that will satisfy the requirements outlined in chapter three then an algorithm cannot be operated.

In chapter five I use the arguments given in chapters three and four to develop transparency regulations that should be followed in order for a criminal justice algorithm to be operated in the New Zealand criminal justice system. Firstly, I consider the kind of regulatory body that New Zealand should have. I argue that whilst a regulatory body will

help to enforce regulations an independent specialist should develop the regulations for the criminal justice system. I look at whether the GDPR, ALGO-CARE, or the 'Right of access by person to reasons for decisions affecting that person' (Official Information Act 1982 (23)) are appropriate ways of ensuring transparency and I argue that none of them meet the requirements that I argued were necessary in chapter three. I then specify the rules that the New Zealand criminal justice system should abide by when operating criminal justice algorithms. I state that whilst these rules surely apply to other jurisdictions these rules are tailored for New Zealand.

Future work and regulating algorithm use

A full ethical assessment of criminal justice algorithms must direct how to build and operate criminal justice algorithms well. My thesis then, acts as a part of this: it explores and directs the ways in which transparency is necessary when using criminal justice algorithms.

Beyond the scope of this thesis there are elements that also need rigorous and comprehensive analysis. Bias and accuracy are commonly regarded as ethical problems for the fair operation of criminal justice algorithms (Kehl et al. 2017). Elements such as these need to undergo the same sort of assessment that transparency has done in this thesis. Developing regulations mandating what to do about bias in, and accuracy of criminal justice algorithms is also necessary.

In concluding this thesis, my hope is that this transparency assessment can serve as a helpful contribution to developing ethical regulations for criminal justice algorithms in New Zealand. This in combination with other ethical assessments can then be used to develop a comprehensive ethical framework that informs regulations for the development and operation of criminal justice algorithms in New Zealand.

Appendices

Appendix A: Rating and level of risk for RoC*RoI

Score	**Level of Risk**
0.0 – 0.2999	Low Risk
0.3 – 0.6999	Medium Risk
0.7 – 0.9999	High Risk

(Johnston 2019)

Appendix B: RoC*RoI factors, descriptions and weightings

Regression Variable	Description/weighting
MaleFirstOffenderFree13	*Log of time not in prison since age 13 for male first time offender*
MaleReoffenderFree13	*Log of time not in prison since age 13 for male re-offender*
FemaleFirstOffenderFree13	*Log of time not in prison since age 13 for female first time offender*
FemaleReoffenderFree13	*Log of time not in prison since age 13 for female re-offender*
MaleEpisodesFree	*Log of time between the two most recent sentence periods ("episodes") of male offenders*
FemaleEpisodesFree	*Log of time between the two most recent sentence periods ("episodes") of female offenders*
Reoffender	*Is this not a first offence – value 1 for yes, 0 for no*
MaleReoffender[1]	*For males, is this not a first offence, for females, is this a first offence*
MaleReoffenderSerious	*History of offending based on seriousness, for males*
FemaleReoffenderSerious	*History of offending based on seriousness, for females*
FemaleFirstOffenderSerious	*Seriousness of offence for first time female offender*
MaleFirstOffenderSerious	*Seriousness of offence for first time male offender*
MaleEpisodeCount	*Total number of sentence periods ("episodes"), for male offender*
FemaleEpisodeCount	*Total number of sentence periods ("episodes"), for female offender*
FemaleOffendingRate	*Number of sentence periods / time not in prison, for females*
MaleOffendingRate	*Number of sentence periods / time not in prison, for males*
PreAge13Offence	*Any offence committed prior to age 13 – value 1 for yes, 0 for no*
Male	*Is offender male – value 1 for yes, 0 for no*
FemaleReoffenderDrive	*Is current most serious offence a traffic offence, and is it not a first offence for female offender – value 1 for yes, 0 for no*
MaleReoffenderDrive	*Is current most serious offence a traffic offence, and is it not a first offence for male offender – value 1 for yes, 0 for no*
FemaleFirstOffenderDrive	*Is current most serious offence a traffic offence, and is this a first offence for female offender – value 1 for yes, 0 for no*
MaleFirstOffenderDrive	*Is current most serious offence a traffic offence, and is this a first offence for male offender – value 1 for yes, 0 for no*

Risk of reimprisonment (RoI):

Regression Variable	Description/weighting
DrugOffender	*Is current most serious offence a drug offence – value 1 for yes, 0 for no*
Male	*Value 1 if offender is male, 0 if not*
MeanSeriousness	*Mean (unweighted) seriousness of all historical convictions*
ConvictionCount	*Log of total number of convictions*
PrisonEpisode	*Value 1 if offender has had prison sentence, 0 if no*
OffendingRate	*Number of sentence periods / time not in prison*
Reoffender	*1 if not first offence, 0 if yes*
TimeFreeEpisodes	*Log of time between 2 most recent sentence periods*
MaleTimeFree13	*Log of time not in prison since age 13 for male offender*
WeightedSeriousness	*History of offending based on seriousness, weighted*

(Johnston 2018)

Appendix C: Sample COMPAS risk assessment questionnaire

Risk Assessment

PERSON			
Name: [redacted]		Offender #: [redacted]	DOB: [redacted]
[redacted]	Gender: Male	Marital Status: Single	Agency: DAI

ASSESSMENT INFORMATION			
Case Identifier: [redacted]	Scale Set: Wisconsin Core - Community Language	Screener: [redacted]	Screening Date: [redacted]

Current Charges

☐ Homicide	☑ Weapons	☑ Assault	☐ Arson
☐ Robbery	☐ Burglary	☐ Property/Larceny	☐ Fraud
☐ Drug Trafficking/Sales	☐ Drug Possession/Use	☐ DUI/OUIL	☑ Other
☐ Sex Offense with Force	☐ Sex Offense w/o Force		

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
   5
8. How many prior juvenile felony offense arrests?
   ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☑ 4 ☐ 5+
9. How many prior juvenile violent felony offense arrests?
   ☐ 0 ☐ 1 ☑ 2+
10. How many prior commitments to a juvenile institution?
    ☐ 0 ☑ 1 ☐ 2+

[redacted] [redacted]

Note to Screener: The following Criminal History Summary questions require you to add up the total number of specific types of offenses in the person's criminal history. Count an offense type if it was among the charges or counts within an arrest event. Exclude the current case for the following questions.

11. How many times has this person been arrested for a felony property offense that included an element of violence?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
12. How many prior murder/voluntary manslaughter offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
13. How many prior felony assault offense arrests (not murder, sex, or domestic violence) as an adult?
☑0 ☐1 ☐2 ☐3+
14. How many prior misdemeanor assault offense arrests (not sex or domestic violence) as an adult?
☑0 ☐1 ☐2 ☐3+
15. How many prior family violence offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
16. How many prior sex offense arrests (with force) as an adult?
☑0 ☐1 ☐2 ☐3+
17. How many prior weapons offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
18. How many prior drug trafficking/sales offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
19. How many prior drug possession/use offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
20. How many times has this person been sentenced to jail for 30 days or more?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
21. How many times has this person been sentenced (new commitment) to state or federal prison?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
22. How many times has this person been sentenced to probation as an adult?
☑0 ☐1 ☐2 ☐3 ☐4 ☐5+

Include the current case for the following question(s).

23. Has this person, while incarcerated in jail or prison, ever received serious or administrative disciplinary infractions for fighting/threatening other inmates or staff?
☑ No ☐ Yes
24. What was the age of this person when he or she was first arrested as an adult or juvenile (criminal arrests only)?
14

Non-Compliance

Include the current case for these questions.

25. How many times has this person violated his or her parole?
☑0 ☐1 ☐2 ☐3 ☐4 ☐5+
26. How many times has this person been returned to custody while on parole?
☑0 ☐1 ☐2 ☐3 ☐4 ☐5+
27. How many times has this person had a new charge/arrest while on probation?
☐0 ☐1 ☐2 ☐3 ☑4 ☐5+
28. How many times has this person's probation been violated or revoked?
☐0 ☐1 ☑2 ☐3 ☐4 ☐5+

[redacted]

[redacted]

29. How many times has this person failed to appear for a scheduled criminal court hearing?
☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+

30. How many times has the person been arrested/charged w/new crime while on pretrial release (includes current)?
☐ 0 ☐ 1 ☐ 2 ☑ 3+

Family Criminality

The next few questions are about the family or caretakers that mainly raised you when growing up.

31. Which of the following best describes who principally raised you?
☐ Both Natural Parents
☐ Natural Mother Only
☐ Natural Father Only
☐ Relative(s)
☐ Adoptive Parent(s)
☐ Foster Parent(s)
☑ Other arrangement

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 ☐ 5 to 10 ☐ 11 to 14 ☐ 15 or older ☐ Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

35. Were your brothers or sisters ever arrested, that you know of?
☐ No ☑ Yes

36. Was your wife/husband/partner ever arrested, that you know of?
☑ No ☐ Yes

37. Did a parent or parent figure who raised you ever have a drug or alcohol problem?
☑ No ☐ Yes

38. Was one of your parents (or parent figure who raised you) ever sent to jail or prison?
☑ No ☐ Yes

Peers

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most

40. How many of your friends/acquaintances served time in jail or prison?
☐ None ☐ Few ☑ Half ☐ Most

41. How many of your friends/acquaintances are gang members?
☐ None ☑ Few ☐ Half ☐ Most

42. How many of your friends/acquaintances are taking illegal drugs regularly (more than a couple times a month)?
☑ None ☐ Few ☐ Half ☐ Most

43. Have you ever been a gang member?
☐ No ☑ Yes

44. Are you now a gang member?
☐ No ☑ Yes

Substance Abuse

What are your usual habits in using alcohol and drugs?

45. Do you think your current/past legal problems are partly because of alcohol or drugs?
☑ No ☐ Yes

46. Were you using alcohol or under the influence when arrested for your current offense?
☐ No ☑ Yes

47. Were you using drugs or under the influence when arrested for your current offense?
☑ No ☐ Yes

48. Are you currently in formal treatment for alcohol or drugs such as counseling, outpatient, inpatient, residential?
☑ No ☐ Yes

49. Have you ever been in formal treatment for alcohol such as counseling, outpatient, inpatient, residential?
☑ No ☐ Yes

50. Have you ever been in formal treatment for drugs such as counseling, outpatient, inpatient, residential?
☑ No ☐ Yes

51. Do you think you would benefit from getting treatment for alcohol?
☑ No ☐ Yes

52. Do you think you would benefit from getting treatment for drugs?
☑ No ☐ Yes

53. Did you use heroin, cocaine, crack or methamphetamines as a juvenile?
☐ No ☑ Yes

Residence/Stability

54. How often do you have contact with your family (may be in person, phone, mail)?
☐ No family ☐ Never ☐ Less than once/month ☐ Once per week ☑ Daily

55. How often have you moved in the last twelve months?
☐ Never ☑ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5+

56. Do you have a regular living situation (an address where you usually stay and can be reached)?
☐ No ☑ Yes

57. How long have you been living at your current address?
☑ 0-5 mo. ☐ 6-11 mo. ☐ 1-3 yrs. ☐ 4-5 yrs. ☐ 6+ yrs.

58. Is there a telephone at this residence (a cell phone is an appropriate alternative)?
☐ No ☑ Yes

59. Can you provide a verifiable residential address?
☐ No ☑ Yes

60. How long have you been living in that community or neighborhood?
☐ 0-2 mo. ☐ 3-5 mo. ☐ 6-11 mo. ☑ 1+ yrs.

61. Do you live with family—natural parents, primary person who raised you, blood relative, spouse, children, or boy/girl friend if living together for more than 1 year?
☐ No ☑ Yes

62. Do you live with friends?
☑ No ☐ Yes

63. Do you live alone?
☑ No ☐ Yes

64. Do you have an alias (do you sometimes call yourself by another name)?
☑ No ☐ Yes

Social Environment

Think of the neighborhood where you lived during the past few (3-6) months.

65. Is there much crime in your neighborhood?
☑ No ☐ Yes

66. Do some of your friends or family feel they must carry a weapon to protect themselves in your neighborhood?
☑ No ☐ Yes

67. In your neighborhood, have some of your friends or family been crime victims?
☐ No ☑ Yes

68. Do some of the people in your neighborhood feel they need to carry a weapon for protection?
☐ No ☑ Yes

69. Is it easy to get drugs in your neighborhood?
☑ No ☐ Yes

70. Are there gangs in your neighborhood?
☐ No ☑ Yes

Education

Think of your school experiences when you were growing up.

71. Did you complete your high school diploma or GED?
☑ No ☐ Yes

72. What was your final grade completed in school?
9

73. What were your usual grades in high school?
☐ A ☐ B ☑ C ☐ D ☐ E/F ☐ Did Not Attend

74. Were you ever suspended or expelled from school?
☐ No ☑ Yes

75. Did you fail or repeat a grade level?
☑ No ☐ Yes

76. How often did you have conflicts with teachers at school?
☐ Never ☑ Sometimes ☐ Often

77. How many times did you skip classes while in school?
☐ Never ☑ Sometimes ☐ Often

78. How strongly do you agree or disagree with the following: I always behaved myself in school?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

79. How often did you get in fights while at school?
☐ Never ☑ Sometimes ☐ Often

Vocation (Work)

Please think of your past work experiences, job experiences, and financial situation.

80. Do you have a job?
☑ No ☐ Yes

81. Do you currently have a skill, trade or profession at which you usually find work?
☑ No ☐ Yes

82. Can you verify your employer or school (if attending)?
☑ No ☐ Yes

83. How much have you worked or been enrolled in school in the last 12 months?
☐ 12 Months Full-time ☐ 12 Months Part-time ☐ 6+ Months Full-time ☑ 0 to 6 Months PT/FT

84. Have you ever been fired from a job?
☑ No ☐ Yes

85. About how many times have you been fired from a job?
0

86. Right now, do you feel you need more training in a new job or career skill?
☑ No ☐ Yes

87. Right now, if you were to get (or have) a good job how would you rate your chance of being successful?
☑ Good ☐ Fair ☐ Poor

88. How often do you have conflicts with friends/family over money?
☐ Often ☐ Sometimes ☑ Never

89. How hard is it for you to find a job ABOVE minimum wage compared to others?
☐ Easier ☑ Same ☐ Harder ☐ Much Harder

90. How often do you have barely enough money to get by?
☐ Often ☑ Sometimes ☐ Never

91. Has anyone accused you of not paying child support?
☑ No ☐ Yes

92. How often do you have trouble paying bills?
☐ Often ☐ Sometimes ☑ Never

93. Do you frequently get jobs that don't pay more than minimum wage?
☐ Often ☐ Sometimes ☑ Never

94. How often do you worry about financial survival?
☐ Often ☐ Sometimes ☑ Never

Leisure/Recreation

Thinking of your leisure time in the past few (3-6) months, how often did you have the following feelings?

95. How often did you feel bored?
☐ Never ☑ Several times/mo ☐ Several times/wk ☐ Daily

96. How often did you feel you have nothing to do in your spare time?
☐ Never ☑ Several times/mo ☐ Several times/wk ☐ Daily

97. How much do you agree or disagree with the following - You feel unhappy at times?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

98. Do you feel discouraged at times?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

99. How much do you agree or disagree with the following -You are often restless and bored?
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

100. Do you often become bored with your usual activities?
☑ No ☐ Yes ☐ Unsure

101. Do you feel that the things you do are boring or dull?
☑ No ☐ Yes ☐ Unsure

102. Is it difficult for you to keep your mind on one thing for a long time?
☑ No ☐ Yes ☐ Unsure

Social Isolation

Think of your social situation with friends, family, and other people in the past few (3-6) months. Did you have many friends or were you more of a loner? How much do you agree or disagree with these statements?

103. "I have friends who help me when I have troubles."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☑ Strongly Agree

104. "I feel lonely."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

[redacted] Version 8.1.18.12 [redacted]

105. "I have friends who enjoy doing things with me."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☑ Strongly Agree

106. "No one really knows me very well."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

107. "I feel very close to some of my friends."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

108. "I often feel left out of things."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

109. "I can find companionship when I want."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

110. "I have a best friend I can talk with about everything."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☑ Strongly Agree

111. "I have never felt sad about things in my life."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

Criminal Personality

The next few statements are about what you are like as a person, what your thoughts are, and how other people see you. There are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

112. "I am seen by others as cold and unfeeling."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

113. "I always practice what I preach."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

114. "The trouble with getting close to people is that they start making demands on you."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

115. "I have the ability to "sweet talk" people to get what I want."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

116. "I have played sick to get out of something."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

117. "I'm really good at talking my way out of problems."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

118. "I have gotten involved in things I later wished I could have gotten out of."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

119. "I feel bad if I break a promise I have made to someone."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

120. "To get ahead in life you must always put yourself first."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

Anger

121. "Some people see me as a violent person."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure ☐ Agree ☐ Strongly Agree

122. "I get into trouble because I do things without thinking."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure ☐ Agree ☐ Strongly Agree

123. "I almost never lose my temper."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

124. "If people make me angry or lose my temper, I can be dangerous."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

[redacted]

[redacted]

125. "I have never intensely disliked anyone."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

126. "I have a short temper and can get angry quickly."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

Criminal Attitudes

The next statements are about your feelings and beliefs about various things. Again, there are no 'right or wrong' answers. Just indicate how much you agree or disagree with each statement.

127. "A hungry person has a right to steal."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

128. "When people get into trouble with the law it's because they have no chance to get a decent job."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

129. "When people do minor offenses or use drugs they don't hurt anyone except themselves."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

130. "If someone insults my friends, family or group they are asking for trouble."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure ☐ Agree ☐ Strongly Agree

131. "When things are stolen from rich people they won't miss the stuff because insurance will cover the loss."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

132. "I have felt very angry at someone or at something."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

133. "Some people must be treated roughly or beaten up just to send them a clear message."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

134. "I won't hesitate to hit or threaten people if they have done something to hurt my friends or family."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure ☐ Agree ☐ Strongly Agree

135. "The law doesn't help average people."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

136. "Many people get into trouble or use drugs because society has given them no education, jobs or future."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

137. "Some people just don't deserve any respect and should be treated like animals."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

[redacted] [redacted]

(Northpointe 2016)

Appendix D: ALGO-CARE framework and explanatory notes

A proposed decision-making framework for the deployment of algorithmic assessment tools in the policing context		
A	**Advisory**	Is the assessment made by the algorithm used in an advisory capacity? Does a human officer retain decision-making discretion? What other decision-making by human officers will add objectivity to the decisions (partly) based on the algorithm?
L	**Lawful**	On a case-by-case basis, what is the policing purpose justifying the use of algorithm, both its means and ends?[a] Is the potential interference with the privacy of individuals necessary and proportionate for legitimate policing purposes? In what way will the tool improve the current system and is this demonstrable? Are the data processed by the algorithm lawfully obtained, processed and retained, according to a genuine necessity with a rational connection to a policing aim? Is the operation of the tool compliant with national guidance?
G	**Granularity**	Does the algorithm make suggestions at a sufficient level of detail/granularity, given the purpose of the algorithm and the nature of the data processed? Is data categorised to avoid 'broad-brush' grouping and results, and therefore issues potential bias? Do the benefits outweigh any technological or data quality uncertainties or gaps? Is the provenance and quality of the data sufficiently sound? Consider how often the data should be refreshed. If the tool takes a precautionary approach towards false negatives, consider the justifications for this.
O	**Ownership**	Who owns the algorithm and the data analysed? Does the force need rights to access, use and amend the source code and data analysed? How will the tool be maintained and updated? Are there any contractual or other restrictions which might limit accountability or evaluation? How is the operation of the algorithm kept secure?

C	**Challengeable**	What are the post-implementation oversight and audit mechanisms e.g. to identify any bias? Where an algorithmic tool informs criminal justice disposals, how are individuals notified of its use (as appropriate in the context of the tool's operation and purpose)?
A	**Accuracy**	Does the specification match the policing aim and decision policy? Can the stated accuracy of the algorithm be validated reasonably periodically? Can the percentage of false positives/negatives be justified? How was this method chosen as opposed to other available methods? What are the consequences of inaccurate forecasts? Does this represent an acceptable risk (in terms of both likelihood and impact)? Is the algorithmic tool deployed by those with appropriate expertise?
R	**Responsible**	Would the operation of the algorithm be considered fair? Is the use of the algorithm transparent (taking account of the context of its use), accountable and placed under review alongside other IT developments in policing? Would it be considered to be for the public interest and ethical?
E	**Explainable**	Is appropriate information available about the decision-making rule(s) and the impact that each factor has on the final score or outcome (in a similar way to a gravity matrix)? Is the force able to access and deploy a data science expert to explain and justify the algorithmic tool (in a similar way to an expert forensic pathologist)?

[a]Or as Brauneis and Goodman put it, what is the 'predictive goal'? n52 (51).

The Algorithms in Policing – Take ALGO-CARE ™ framework is intended to provide guidance for the use of risk-assessment, predictive, forecasting, classification, decision-making and assistive policing tools which incorporate algorithmic machine learning methods and which may impact individuals on a micro or macro level		
A	**Advisory**	Care should be taken to ensure that an algorithm is not inappropriately fettering an officer's discretion, as natural justice and procedural fairness claims may well arise. Consider if supposedly advisory algorithmic assessments are in practice having undue influence. If it is proposed that an algorithmic decision be automated and determinative, is this justified by the factors below? Data protection rights in regard to automated decisions may then apply.
L	**Lawful**	The algorithm's proposed functions, application, individual effect and use of datasets (police-held data and third party data) should be considered against necessity, proportionality and data minimisation principles, in order to inform a 'go/no-go' decision. In relation to tools that may inform criminal justice disposals, regard should be given to the duty to give reasons.
G	**Granularity**	Consideration should be given to common problems in data analysis, such as those relating to the meaning of data, compatibility of data from disparate sources, missing data and inferencing. Do forces know how much averaging or blurring has already been applied to inputs (e.g. postcode area averages)?
O	**Ownership**	Consider intellectual property ownership, maintenance of the tool and whether open source algorithms should be the default.[a] When drafting procurement contracts with third party software suppliers (commercial or academic), require disclosure of the algorithmic workings in a way that would facilitate investigation by a third party in an adversarial context if necessary. Ensure the force has appropriate rights to use, amend and disclose the tool and any third party data. Require the supplier to provide an 'expert' witness/evidence of the tool's operation if required by the force.[b]

C	**Challengeable**	The results of the analysis should be applied in the context of appropriate professional codes and regulations. Consider whether the application of the algorithm requires information to be given to the individual and/or legal advisor. Regular validation and recalibration of the system should be based on publicly observable (unless non-disclosable for policing/national security reasons) scoring rules.
A	**Accuracy**	How are results checked for accuracy, and how is historic accuracy fed back into the algorithm for the future? Can forces understand how inaccurate or out-of-date input data affects the result?
R	**Responsible**	It is recommended that ethical considerations, such as consideration of the public good and moral principles (so spanning wider concerns than legal compliance) are factored into the deployment decision-making process. Administrative arrangements such as an ethical review committee incorporating independent members could be established for such a purpose (such as Cleveland & Durham Joint External Ethics Committee[c] or the National Statistician's Data Ethics Advisory Committee).[d]
E	**Explainable**	The latest methods of interpretable and accountable machine learning systems should be considered and incorporated into the specification as appropriate.[e] This is particularly important if considering deployment of 'black box' algorithms, where inputs and outputs are viewable but internal workings are opaque (the rule emerges from the data analysis undertaken). Has the relevant Policing & Crime Commissioner been briefed appropriately?

(Oswald et al. 2018, 245–48)

References

Alpaydin, Ethem. 2016. *Machine Learning*. Cambridge: MIT Press.

Andrews, Leighton, Bilel Benzoubid, Lee A. Bygrave, David Demortain, Alex Griffiths, Martin Lodge, Andrea Mennicken, and Karen Yeung. 2017. *Algorithmic Regulation*. London: London School of Economics and Political Science. https://www.kcl.ac.uk/law/research/centres/telos/assets/DP85-Algorithmic-Regulation-Sep-2017.pdf.

Australian Government Department of Health. 2019. "Predictive Risk Model Algorithm." Australian Government Department of Health. March 21. http://www.health.gov.au/internet/main/publishing.nsf/Content/predictive-risk-model-algorithm (accessed 28/03/2019).

Bakker, L. W, David Riley, James O'Malley. 1999. *Risk of Reconviction: Statistical Models Predicting Four Types of Re-Offending*. Wellington: Department of Corrections.

Barker, Anthony. 2001. "Accountability and Responsibility of Government and Public Bodies." *The Political Quarterly*, 72 (1): 132-140.

Barnes, Geoffrey C., and Jordan M. Hyatt. 2012. *Classifying Adult Probationers by Forecasting Future Offending: Final Technical Report*. Philadelphia: Jerry Lee Center of Criminology, University of Pennsylvania. https://www.ncjrs.gov/pdffiles1/nij/grants/238082.pdf

Baxt, W.G. 1995. "Applications of Artificial Neural Networks to Clinical Medicine." *Lancet,* 346 (8983): 1135-1138.

Beech, Anthony, Caroline Friendship, Matt Erikson, and R. Karl Hanson. 2002. "The Relationship Between Static and Dynamic Risk Factors and Reconviction in a Sample of U.L. Child Abusers." *Sexual Abuse: A Journal of Research and Treatment*, 14(2): 155-167.

Bennet Moses, Lyria, and Janet Chan. 2018. "Using Big Data and Data Analytics in Criminological Research" in *Doing Criminological Research*, edited by Pamela Davies and Peter Frances, 3rd ed. 251-70. London: Sage Publications Ltd.

Bird, Francis, W. 1913. "The Evolution of Due Process of Law in Decisions of the United States Supreme Court." *Columbia Law Review*, 12(1): 37-50.

Bivins, Thomas. 2006. "Responsibility and Accountability" In *Ethics in Public Relations: Responsible Advocacy*, edited by Kathy Fitzpatrick and Carolyn Bronstein, 19-38. Sage Publications Ltd.

Bloom, Paul. 2016. *Against Empathy: The Case for Rational Compassion*. New York: Ecco.

Bousquet, Chris and Stephen Goldsmith. 2018. "The Right Way to Regulate Algorithms" Harvard University Data Smart-City Solutions. April 3. https://datasmart.ash.harvard.edu/news/article/right-way-regulate-algorithms (accessed 26/02/2019).

Brennan Tim, Willian Dieterich, and Beate Ehret. 2009. "Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System." *Criminal Justice and Behavior*, 36(1): 21-40.

California Coalition on Sexual Offending. 2015. "Risk Assessment: A Short Introduction-Part 1*"*. California: Californian Coalition on Sexual Offending. https://ccoso.org/sites/default/files/AssessmentPrimer1.pdf (acccessed 06/06/2018).

Caplan, Robyn. 2018. *Algorithmic Accountability: A Primer*. Washington DC: Data & Society. https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf.

Casey, Pamela M., Jennifer L. Elek, Katherine Holt, Tracey D. Johnson, Shelley Spacek Miller, and Roger L. Warren. 2014. *Use of Risk and Needs Assessment Information at Sentencing: La Crosse Country, Wisconsin*. Williamsburg: Centre for Sentencing Initiatives, National Center for State Courts. https://www.ncsc.org/~/media/Microsites/Files/CSI/RNA%20Brief%20%20La%20Crosse%20County%20WI%20csi.ash.

Chen, Li M. 2015. "Overview of Basic Methods for Data Science" In *Mathematical Problems in Data Science*, by Li M. Chen, Zhixun Su, and Bo Jiang, 17-37. Cham: Springer International Publishing.

Coffin, Marie, and Matthew J. Saltzman. 2000. "Statistical Analysis of Computational Tests of Algorithms and Heuristics." *INFORMS Journal on Computing,* 12(1):24-44

Craig, Leam, and Anthony Beech. 2009. "Best Practice in Conducting Actuarial Risk Assessments with Adult Sexual Offenders." *Journal of Sexual Aggression,* 12(2): 193-211.

Crimes Act 1961. (New Zealand) No.43. http://www.legislation.govt.nz/act/public/1961/0043/latest/whole.html.

Criminal Procedure Act 2011. (New Zealand). No. 81. http://www.legislation.govt.nz/act/public/2011/0081/153.0/DLM3359962.html.

Curtis, Jacob. 2018. "On Using Machine Learning to Predict Recidivism," PhD diss., Texas Tech University. https://ttu-ir.tdl.org/bitstream/handle/2346/…/CURTIS-DISSERTATION-2018.pdf?

Dare, Tim and Eileen Gambrill. 2017. "Ethical Analysis: Predictive Risk Models at Call Screening" in *Developing Risk Models to Support Child Maltreatment Hotline Screening Decisions*. Centre for Social Data Analytics. Auckland: AUT.

Desmarais, Sarah, and Jay Singh. 2013. *Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States*. Council of State Governments Justice Center. https://csgjusticecenter.org/wp-content/uploads/2014/07/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf.

Department of Corrections.1999. *Risk of Reconviction*. Wellington: Department of Corrections. https://www.corrections.govt.nz/resources/research_and_statistics/risk-of-reconviction.html.

Dieterich, William, Christina Mendoza, and Tim Brennan. 2016. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity: Performance of the COMPAS Risk Scales in Broward County" *Northpoint Inc Research Department*. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data**,** 1995 O.J. (L281)

Downie, R.S., and Elizabeth Telfer. 1971. "Autonomy." *The Royal Institute of Philosophy,* 46 (178): 293–301.

Dressel, Julia, and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances,* 4 (1): 1-5.

Eaglin, Jessica M. 2017. "Constructing Recidivism Risk." *Emory Law Journal*, 67(59): 59-122.

Edwards, Lilian and Michael Veale. 2017. “Slave to the Algorithm? Why a ‘Right to Explanation’ Is Probably Not the Remedy You Are Looking For.” *Duke Law and Technology Review,* 16 (1): 18-84.
EPFL IRGC. 2018. *The Governance of Decision-Making Algorithms*. Lausanne: EPFL International Risk Governance Centre. https://irgc.epfl.ch/wp-content/uploads/2018/11/IRGC-2018-The-Governance-of-Decision-Making-Algorithms-Workshop-report.pdf.
Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: St Martins Press.
European Commission. 2018. “Data Protection in the European Union.” European Commission. (accessed 03/03/2019). https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en.
Eyal, Nir. 2019. “Informed Consent.” *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. http://plato.stanford.edu/archives/fall2012/entries/informed-consent/.
Freeman, Katherine. 2016. “Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis.” *North Carolina Journal of Law & Technology,* 18 (5): 75–106.
Freeman- Brown, Jane. 2013. “Why keep offender’s secrets? The pros and cons of confidentiality.” *Practice – The New Zealand Corrections Journal*, 1(1): 18-19.
Fry, Hannah. 2018. *Hello World: Being Human in the Age of Algorithms*. London: Doubleday.
General Data Protection Regulation 2018. https://gdpr-info.eu/.
Gendreau, Paul, Yvette Thériault, Paula Smith, and Myrinda Schweitzer. 2012. *Appendix 15: Review of the Rehabilitation and Reintegration Services (RRS) For the New Zealand Department of Corrections*. Wellington: Parliament NZ. https://www.parliament.nz/resource/0000240839.
Gibney, Elizabeth. 2016. “Google AI Algorithm Masters Ancient Game of Go.” *Nature,* 529 (7587): 445–46.
Goodman, Bryce, and Seth Flaxman. 2017. “EU Regulations on Algorithmic Decision-Making and a ‘Right to Explanation.’” *AI Magazine,* 38 (3): 26–30.
Gottfredson, Stephen D. and Laura Moriarty. 2006. “Statistical Risk Assessment: Old
Problems and New Applications.” Crime & Delinquency, 52(1): 178-200.
Gunkel, David J. 2018. *Gaming the System: Deconstructing Video Games, Game Studies, and Virtual Worlds*. Bloomington, Indiana: Indiana University Press.
Guthrie Fergusson, Andrew. 2017. *The Rise of Big Data Policing*. New York: NYU Press.
Hanson, R. Karl, Andrew J. R. Harris, Terri-Lynne Scott, and Leslie Helmus. 2007. *Assessing the Risk of Sexual Offenders on Community Supervision: The Dynamic Supervision Project*. Ottawa: Public Safety Canada. https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ssssng-rsk-sxl-ffndrs/ssssng-rsk-sxl-ffndrs-eng.pdf.
Heald, David. 2006a. “Transparency as an Instrumental Value.” In *Transparency: The Key to Better Governance?*, edited by Christopher Hood and David Heald, 59–73. Oxford: Oxford University Press.
———. 2006b. “Varieties of Transparency.” In *Transparency: The Key to Better Governance*, edited by Christopher Hood and David Heald, 25–43. Oxford,: Oxford University Press.
Hettinger, Edwin C.1989. “*Justifying Intellectual Property*.” Philosophy & Public Affairs, 18 (1): 31-52.

Hoffman, Martin. 2000. *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge, United Kingdom: Cambridge University Press.
Hood, Christopher. 2010. "Accountability and Transparency: Siamese Twins, Matching Parts, Awkward Couple?" *West European Politics,* 33(5): 989–1009.
Hope, Bruce, K. 2007. What's Wrong with Risk Assessment?." 13(6): 1159-1163.
Johnston, Peter. 2018. "Research on RoC*RoI and Transparency," December 5, 2018.
———. 2019. "Research on RoC*RoI," March 13, 2019.
Judicial Review Procedure Act 2016. No. 50. http://www.legislation.govt.nz/act/public/2016/0050/latest/whole.html#DLM6942108.
Kahneman, Daniel. 2011. *Thinking Fast and Slow*. New York, NY: Farrar, Straus and Giroux.
Kehl, Danielle, Priscilla Guo, and Sam Kessler. 2017. "Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing" Responsive Communities Initiative, Harvard: Berkman Klein Center, Harvard Law School.
Kim, Phil. 2017. *MATLAB Deep Learning: With Machine Learning, Neural Networks, and Artificial Intelligence*. Berkeley: APress Berkeley.
Kroll, Joshua A., Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G Robinson, and Harlan Yu. 2017. "Accountable Algorithms." *University of Pennsylvania Law Review,* 165: 633–705.
Larson, Jeff, Surya Mattu, Lauren Kirchner, and Angwin Julia. "How We Analysed the COMPAS Recidivism Algorithm." ProPublica. May 23, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed on 10/04/2018).
LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature,* 521 (7553): 436–44.
Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* Boston, MA: Houghton Mifflin Harcourt.
McCaney, Kevin. "Where Do Humans Outperform AI?" Government CIO Media. April 4, 2018. https://www.governmentciomedia.com/where-do-humans-outperform-ai (accessed 13/12/2018).
McDonald, John H. 2014. "Multiple Logistic Regression." *Handbook of Biological Statistics*. Baltimore Maryland: Sparky House Publishing. http://www.biostathandbook.com/multiplelogistic.html.
Michigan Department of Corrections. 2017. Administration and Use of COMPAS in the Presentence Investigation Report. Lansing: Michigan Department of Corrections. https://www.michbar.org/file/news/releases/archives17/COMPAS-at-PSI-Manual-2-27-17-Combined.pdf.
Ministry of Business, Innovation and Employment. "Types of Intellectual Property." Ministry of Business, Innovation and Employment. 2019. https://www.business.govt.nz/risks-and-operations/intellectual-property-protection/types-of-intellectual-property/ (accessed 04/03/2019).
Mittelstadt, Brent Daniel, Patrick Allo, Mariarosari Taddeo, Sandra Wachter, Luciano Floridi. 2016. "The ethics of algorithms: Mapping the debate". *Big Data and Society*, July-December: 1-21.
Moore, Adam, and Ken Himma. 2014. "Intellectual Property." *Stanford Encylopedia of Philosophy*. Stanford: University of Stanford.
https://plato.stanford.edu/entries/intellectual-property/.
Moore, Taylor R. 2017. *Trade Secrets & Algorithms as Barriers to Social Justice*. Washington DC: Centre for Democracy and Technology.

https://cdt.org/files/2017/08/2017-07-31-Trade-Secret-Algorithms-as-Barriers-to-Social-Justice.pdf.
Mulgan, Richard. 2000. "'Accountability': An Ever Expanding Concept?" *Public Administration,* 78 (3): 555–73.
Nadesu, Arul. 2007. *Reconviction Patterns of Released Prisoners: A 36-Months Follow-up Analysis*. Wellington: Department of Corrections. https://www.corrections.govt.nz/__data/assets/pdf_file/0004/672061/reimprisonment-report.pdf
National Institute of Justice. 2014. "Recidivism". National Institute of Justice. June 17 2014. https://www.nij.gov/topics/corrections/recidivism/Pages/welcome.aspx (accessed on 13/12/2018).
New, Joshua and Daniel Castro. 2018. *How Policymakers Can Foster Algorithmic Accountability*. Washington D.C: Centre for Data Innovation. http://www2.datainnovation.org/2018-algorithmic-accountability.pdf
Northpointe. 2015. "Practitioners Guide to COMPAS." Northpointe. http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf
Northpointe. "COMPAS Risk Assessment." Unpublished document, February 1 2016. PDF File. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html (accessed 24/07/2018).
New Zealand Intellectual Property Office. 2018/ "Patentable Ideas". New Zealand Intellectual Property Office. 2018. https://www.iponz.govt.nz/about-ip/patents/ (accessed 03/03/2019).
New Zealand Transport Agency. "Alcohol and Drug Limits" New Zealand Transport Agency. November 2014 (https://www.nzta.govt.nz/resources/roadcode/motorcycle-road-code/about-limits/alcohol-and-drugs-limits/ (accessed 03/03/2019).
Official Information Act 1982. No. 156. legislation.govt.nz/act/public/1982/0156/107.0/whole.html#DLM65628.
O'Neil, Cathy. 2016. *Weapons of Math Destruction*. New York: Crown Publishers
O'Neill, O. 2003. "Some Limits of Informed Consent." *Journal of Medical Ethics,* 29 (1): 4–7.
O'Neill, O. 2004. "Accountability, Trust and Informed Consent in Medical Practice and Research." *Clinical Medicine,* 4 (3): 269–76.
Oswald, Marion, Jamie Grace, Sheena Urwin, and Geoffrey C. Barnes. 2018. "Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality." *Information & Communications Technology Law,* 27 (2): 223–50.
Parole Act 2002. No. 10. http://www.legislation.govt.nz/act/public/2002/0010/85.0/DLM139306.html
Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.
Public Law Project. "A Brief Guide to Judicial Review" Public Law Project. 2006. https://publiclawproject.org.uk/wp-content/uploads/data/resources/114/PLP_2006_Guide_To_JR_Procedure.pdf
Pullar-Strecker, Tom. 2012. "Big Data Is Watching You." Stuff. August 22, 2012. http://www.stuff.co.nz/technology/digital-living/7501502/Big-Data-is-watching-you.
R v Peta [2007]. NZCA 28.
Rampell, Catherine. "Gaming the System." New York Times. February 14, 2013. https://economix.blogs.nytimes.com/2013/02/14/gaming-the-system/ (accessed 23/01/2019).

Resnik, D.B. 2003. “A Pluralistic Account of Intellectual Property.” *Journal of Business Ethics,* 46: 319–35.

Ribiero, Marco Tulio, Sameer Singh, Carlos Guestrin. 2016. " "Why should I trust you?" Explaining the predictions of any classifier." In *KDD Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Minings.* San Francisco,2016, 1135-1144.

Rodriguez Ferrere, M B. 2012. “The Unnecessary Confusion in New Zealand’s Appellate Jurisdiction.” *Otago Law Review,* 12 (4): 829–40.

Rudin, Cynthia. 2018. “Please Stop Explaining Black Box Models for High-Stakes Decisions” In NIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning. Palais des Congrès de Montréal, Montréal, 2018, 1-15.

Rudin, Cynthia, Caroline Wang, Beau Coker. 2018. “The age of secrecy and unfairness in recidivism.” Draft on arXiv. 1-38. https://arxiv.org/pdf/1811.00731.pdf.

Segal, Michael. “We Need an FDA for Algorithms.” Nautilus, November 2018. http://nautil.us/issue/66/clockwork/we-need-an-fda-for-algorithms (accessed 12/12/2018).

Select Committee on Artificial Intelligence. 2018. *AI in the UK; Ready, Willing and Able?* London: House of Lords. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Senior Court Act 2016. No.48. http://www.legislation.govt.nz/act/public/2016/0048/latest/whole.html?search=sw_096be8ed8180d2b0_appeal_25_se&p=1#DLM5759425

Srivastava, Tavish. “Difference between Machine Learning and Statistical Modeling.” Analytics Vidhya. July 1, 2015. https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/ (accessed 11/6/2018).

State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. 2017. *Harvard Law Review* 130(5):1530–37. https://harvardlawreview.org/2017/03/state-v-loomis/.

Stats NZ. 2018. *Algorithm Assessment Report*. Wellington: Stats NZ. https://www.data.govt.nz/assets/Uploads/Algorithm-Assessment-Report-Oct-2018.pdf.

Steidman, John C. 1962. “Trade Secrets.” *Ohio State Law Journal,* 4: 4–34.

Stergiou, Christos, and Dimitrios Siganos. 2011. *Neural Networks.* London: Imperial College. https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.

Tadros, Victor. 2011 "Consent to Harm" *Current Legal Problems*, 64:23-49.

Tännsjö, Torbjörn. 2014. “Utilitarianism and Informed Consent.” *Journal of Medical Ethics* 40 (7): 445–445.

Teshnizi, Saeed, and Sayyed Ayatollahi. 2015. “A Comparison of Logistic Regression Model and Artificial Neural Networks in Predicting of Student’s Academic Failure.” *Acta Informatica Medica,* 23(5): 296.

Tosun, Erdi, Kadir Aydin, and Mehmet Bilgili. 2016. “Comparison of Linear Regression and Artificial Neural Network Model of a Diesel Engine Fueled with Biodiesel-Alcohol Mixtures.” *Alexandria Engineering Journal,* 55 (4): 3081–89.

Tutt, Andrew. 2017. “An FDA for Algorithms.” *Administrative Law Review,* 69(1): 83–123.

Van der Meer, Tom W.G.. 2018. “Political Trust and the “Crisis of Democracy".” *Oxford Research Encyclopedia of Politics*. Oxford University Press.

Vethan Law Firm. “Trade Secrets: 10 of the Most Famous Examples”. Vethan Law Firm. February 11 2016. https://info.vethanlaw.com/blog/trade-secrets-10-of-the-most-famous-examples (accessed 30/11/2018).

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.” *International Data Privacy Law,* 7 (2): 76–99.

Waitangi Tribunal. 2005. *The Offender Assessment Offender Policies Report*. Wellington: Legislation Direct. https://forms.justice.govt.nz/search/Documents/WT/wt_DOC_68001752/Offender%20Assessment%20Policies.pdf

Wanna, John "Opening government in the information age" Wanna, John, and Sam Vincent, eds. 2018. *Opening Government: Transparency and Engagement in the Information Age*. Canberra: ANU Press.

Ward, Tony, and Clare-Ann Fortune. 2016. “The Role of Dynamic Risk Factors in the Explanation of Offending.” *Aggression and Violent,* 29: 79–88.

Weber, Dane Joseph. 2002. “A Critique of Intellectual Property Rights.” Virginia: Christendom College. http://dane.weber.org/concept/thesis.html#criti.

Wexler, Rebecca. 2018. “Life, Liberty, and Trade Secrets.” *Stanford Law Review,* 70 (5): 1343–1429.

Zarsky, Tal. 2013. “Transparency in Data Mining: From Theory to Data Practice”. In *Discrimination and Privacy in the Information Society*, edited by Bart Custers, Toon Calders, and Bart Schermer, 3:301-24. Heidelberg: Springer.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2018. “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?” *Philosophy & Technology*, September:1-23.